A. N. Matveev

Optics

Mir Publishers Moscow

Translated from Russian by Ram Wadhwa

First published 1988
Revised from the 1985 Russian edition

На английском языке

Printed in the Union of Soviet Socialist Republics

ISBN 5-03-001133-1

© Издательство «Высшая школа», 1985
© English translation, Mir Publishers, 1988

Preface

A rapid introduction of the achievements of science in industry is the distinguishing feature of modern life. Optics is just one striking example of this trend. Intensive development of laser technology has long ceased to be just a landmark in scientific research and has revolutionalized the industrial technology in many fields. Training of qualified personnel for those branches of industry where the latest achievements of science are implemented has become one of the most important problems of our time.

The application of lasers and their use in conjunction with computers have created quite favourable conditions for the development of optics. The high coherence of laser radiation permits an investigation and reproduction of a wide range of phenomena in the optical range. These phenomena cannot be studied without a highly coherent radiation. The high energy density of laser radiation opens new horizons for the investigation of nonlinear processes in optics under conditions that were hitherto unattainable. The possibility of generating short and ultrashort laser pulses has paved the way for analyzing rapid processes including intramolecular ones. The use of computers has given a tremendous impetus to optical studies by reducing them to direct computations or to numerical experiments.

All this has led to a significant advancement of optics in the last quarter of a century and has considerably widened the scope of its applications. This process was triggered by important studies leading to the creation of quantum-mechanical oscillators, viz. lasers and masers. Soviet physicists have not only carried out fundamental research in the fields of lasers and masers, but also made significant contributions towards the development of many important branches of optics, including light scattering, holography, optical systems, nonlinear optics, etc. Naturally, the basic concepts of optics have not undergone significant changes in the course of its recent development. In some cases, these concepts were clarified, while in some other cases they were enriched by the introduction of concepts, methods and mathematical approaches from other branches of science (like the theory of random processes, physics of linear and nonlinear oscillations, matrix computational methods, etc.).

The subject matter of the book is completely reflected in Contents. Statistical properties of light and its spectral representation are covered in greater detail than usual. Diffraction of light is described in the framework of Kirchhoff's integral. The effectiveness of the matrix methods is shown in sections covering geometrical optics and interference of light in thin films. A unified approach involving Fourier optics has been adopted for describing the diffraction theory of image formation, spatial filtration of images, holography, and other allied topics. Analysis of partial coherence and partial polarization is carried out in terms of the first correlation function.

The mathematical aspect of the material presented in this book has been kept as simple as possible, and at the same time in line with the rigorous scientific approach. Wherever necessary, mathematical elucidations are presented and more detailed computations are made. The cumbersomeness of some of these presentations should not produce the impression that the mathematical apparatus used in the book is complicated. The student should muster patience and carry out these calculations independently to rid himself of such an impression.

The most significant aspect in which this book differs from the books dealing with mechanics, molecular physics and electricity is that its basic principles lie beyond the scope of this course. Because of this, considerable emphasis has been laid on the deductive method of description. The material of this course is therefore presented in deductive form and in most cases (though not always) the experimental results are analyzed to show the agreement between the theoretical results and the experimental data, or to explain the observed phenomena.

The book is based on the author's experience of teaching physics for many years at the Physics Faculty of the Lomonosov State University, Moscow. The author is grateful to his colleagues for many fruitful discussions on the content of the book.

The author is indebted to Academician A. I. Akhiezer of the Academy of Sciences of the Ukrainian SSR, as well as to Prof. N. I. Kaliteevskii and his associates at the Department of General Physics, Zhdanov State University, Leningrad, for carefully reviewing the manuscript and making valuable comments.

Moscow, May 1987.

Author

Contents

Preface

1. Electromagnetic Waves

Sec. 1. **Optical Range of Electromagnetic Waves** 15
Wavelengths of the visible range. Wave frequencies in the visible range. Optical and other ranges of electromagnetic waves. Why can we see only in the visible range? Why is microwave range unsuitable for vision? Night vision.

Sec. 2. **Properties of Electromagnetic Waves** 22
Electromagnetic nature of light. Wave equation. Plane waves. Spherical waves. Plane harmonic waves. Wave vector. Complex representation of a plane wave. Complex representation of a spherical wave. Plane electromagnetic wave. Invariance of a plane wave. Phase invariance. Four-dimensional wave vector. Transformation formulas for frequency and direction of propagation of a plane wave. Doppler effect.

Sec. 3. **Flux Densities of Energy and Momentum of Electromagnetic Waves. Light Pressure** 35
Energy flux density. Flux density distribution over beam cross section. Gaussian beam. Momentum density of electromagnetic waves. Light pressure. Effect of light pressure on small particles. Laser fusion. Transformation of amplitude and normal vector of a plane electromagnetic wave. Energy of a plane wave train. Momentum of a plane wave train.

Sec. 4. **Superposition of Electromagnetic Waves** 43
Superposition of field vectors of a wave. Superposition of travelling monochromatic electromagnetic plane waves. Beats. Standing waves. Energy transformation in a standing electromagnetic wave. Experimental proof of the electromagnetic nature of light.

Sec. 5. **Polarization of Electromagnetic Waves** 49
Polarization. Linear polarization. Superposition of linearly polarized waves. Elliptical and circular polarizations. Variation of the electric field vector in space for elliptical and circular polarizations. Degenerate case of elliptical polarization. Number of independent polarizations. Linearly polarized wave as a superposition of circularly polarized waves.

Sec. 6. **Averaging** 52
Averaging operation. Averaging of harmonic functions. Averaging of squares of harmonic functions. Linearity of the averaging operation. Calculations involving complex scalars. Calculations involving complex vector quantities.

Sec. 7. **Photometric Concepts and Quantities** 57
Radiant and photometric quantities. Radiant quantities. Radiant intensity. Radiance. Radiant emittance (exitance). Radiant illuminance. Photometric quantities. Luminous flux. Luminance. Luminous emittance. Illuminance. Light exposure. Relations between radiant and luminous characteristics of radiation.

2. Nonmonochromatic and Random Radiation

Sec. 8. **Spectral Composition of Functions 71**
Fourier series in real form. Fourier series in complex form. Fourier integral in real form. Fourier integral in complex form. Amplitude spectrum and phase spectrum. Determination of the amplitude and phase spectra from complex form of Fourier series. Continuous spectrum. Spectrum of rectangular pulses. Spectrum of sawtooth pulses. Spectrum of a single rectangular pulse. Spectrum of an exponentially decreasing function. Relation between the pulse duration and the spectral width. Displacement of the time reference point. Frequency displacement of spectrum. Negative frequencies. Parseval's theorem. Plancherel's theorem.

Sec. 9. **Natural Linewidth of Radiation 79**
Classical model of a radiator. Spectral composition of radiation. Lorentz shape and width of emission line. Emission time. Shape of absorption line. Quantum interpretation of the shape of emission line. Quasi-monochromatic wave.

Sec. 10. **Spectral Line Broadening 86**
Reasons for broadening. Uniform and nonuniform broadening. Natural width of emission line as a uniform broadening. Collision broadening. Doppler broadening. Shape of a composite emission line.

Sec. 11. **Modulated Waves 91**
Modulation. Amplitude modulation. Frequency and phase modulation. Oscillation spectrum with harmonic frequency modulation.

Sec. 12. **Wave Packets 94**
Wave packet formed by two waves. Group velocity. Superposition of oscillations with equidistant frequencies. Quasi-plane wave.

Sec. 13. **Random Light 98**
Superposition of waves with random phases. Resolution time. Averaging over oscillation period. Effect of increasing time interval on the result of averaging. Coherence time. Coherence length. Gaussian light. Energy flux density fluctuations for random light. Polarization.

Sec. 14. **Fourier Analysis of Random Processes 103**
Power spectrum (spectral function). Autocorrelation function. Wiener-Khintchine theorem. Correlation interval. Relation between the correlation interval and the normalized spectral function.

3. Propagation of Light in Isotropic Media

Sec. 15. **Propagation of Light in Dielectrics 111**
Monochromatic waves. Dispersion. Normal dispersion. Anomalous dispersion. Scattering of light. Propagation of a wave packet. Replacement of a light wave in a medium. Dispersion of light in interstellar space. Colour of bodies.

Sec. 16. **Reflection and Refraction of Light at the Interface Between Two Dielectrics. Fresnel's Formulas 121**
Boundary conditions. Constancy of the wave frequency upon reflection and refraction. The plane of incident, reflected and refracted rays. Relation between the angles of incidence, reflection and refraction. Decomposition of a plane wave into two waves with mutually perpendicular linear polarizations. Vector E is perpendicular to the plane of incidence. Fresnel's formulas for the perpendicular components of the field vector. Vector E lies in the plane of incidence. Fresnel's formulas for parallel components of field vector. Brewster effect. Relation between the phases of reflected and refracted waves. Degrees of polarization.

Sec. 17. **Total Reflection of Light 135**
Formulas for angles $\theta_{in} \geqslant \theta_{lim}$. Wave in the second medium. Penetration depth. Phase velocity. Reflected wave.

Sec. 18. **Energy Relations for Reflection and Refraction of Light 139**
Energy flux densities. Reflection coefficient. Transmission coefficient. Law of energy conservation. Polarization of light upon reflection and refraction.

Sec. 19. **Propagation of Light in Conducting Media 142**
Complex permittivity. Penetration depth. Physical reason behind absorption. Phase velocity and wavelength. Relation between the phases of field vector oscillations. Relation between amplitudes of field vectors. Media with low conductivity. Media with high conductivity.

Sec. 20. **Reflection of Light from a Conducting Surface 147**
Boundary conditions. Relation between wave amplitudes. Reflection coefficient. Connection between reflecting and absorbing properties.

4. Geometrical Optics

Sec. 21. **Geometrical Optics Approximation 153**
Eikonal equation. Ray of light. Range of applicability of the ray approximation. Fermat's principle. Derivation of the law of refraction from Fermat's principle. Propagation of a ray in a medium with varying refractive index.

Sec. 22. **Lenses, Mirrors and Optical Systems 159**
Paraxial approximation. Refraction at a spherical surface. Matrix notation. Propagation of a ray in a lens. Refraction of the ray at the second spherical surface. Refraction of a ray by a lens. Propagation of a ray through an optical system. Reflection from spherical surfaces.

Sec. 23. **Optical Image 165**
Matrix of an optical system. Transformation of a ray from object plane to image plane. Cardinal elements of an optical system. Physical meaning of the Gauss constants. Construction of images. Lens equation. Thin lenses. A system of thin lenses. Utilization of computers.

Sec. 24. **Optical Aberration 173**
Sources of aberration. Exact transformation matrices. Spherical aberration. Coma. Aberration caused by off-axis inclined rays. Chromatic aberration. Oil-immersion lens. Abbe's sine condition.

Sec. 25. **Optical Instruments 180**
Diaphragming. Basic concepts associated with diaphragming. Eye as an optical system. Photographic camera. Magnifying glass. Microscope. Telescope. Optical projectors.

5. Interference

Sec. 26. **Two-beam Interference Caused by Amplitude Division 189**
Definition of interference. Light intensity upon superposition of two monochromatic waves. Methods of producing coherent waves in optics. Interference of monochromatic waves propagating strictly along the axis of a Michelson interferometer. Interference of monochromatic waves propagating at an angle to the interferometer axis. The reason behind the blurring of interference fringes. Interference of nonmonochromatic light. Fourier transform spectroscopy. Visibility for a Gaussian line. Visibility for a Lorentz line. Michelson's interferometer with linear fringes. White light interference pattern. Mach-Zehnder interferometer. Twyman-Green interferometer. Jamin refractometer.

Sec. 27. **Two-beam Interference Through Wave Front Splitting 205**
Huygens' principle. Young's two-slit interferometer. White light interference.

Finite-sized source. Source with a uniform distribution of emission intensity. Space and time coherence. Coherence angle and coherence width. Stellar interferometer. Measurement of the diameters of stars. Measurement of the distance between components of a double star. Fresnel biprism. Billet split lens. Lloyd's mirror interference. Fresnel mirrors. Law of energy conservation in interference.

Sec. 28. **Multiple-beam Interference Through Amplitude Division 216**
Fabry-Perot interferometer. Intensity distribution in interference pattern. Interference rings. Resolving power. Factors limiting the resolving power. Dispersion region. Fabry-Perot scanning interferometer. Interference filters. Lummer-Gehrcke plate. Michelson echelon grating.

Sec. 29. **Interference in Thin Films 226**
Optical path length. Reflection from parallel surfaces. Uniform inclination fringes. The role of source size. The role of plate thickness and monochromaticity of radiation. Uniform thickness fringes (isopachic fringes). Newton rings. Multiple reflection. Layer with zero reflectivity. Layer with a high reflectivity. Matrix method of computations for multilayer films. Multilayer dielectric mirrors. Translucent materials.

Sec. 30. **Partial Coherence and Partial Polarization 239**
Partial coherence. Mutual coherence function. Normalized coherence function. Coherence function. Brown-Twiss experiment. Partial polarization. Coherence matrix for a quasi-monochromatic plane wave. Normalized coherence function for mutually perpendicular projections of the electric field strength of a wave. Natural (nonpolarized) light. Completely polarized light. Degree of polarization of a light wave. Expression of degree of polarization in terms of extremal intensities. Representation of natural light. Relation between the degrees of polarization and coherence function. Van-Zittert-Zernike theorem.

6. Diffraction

Sec. 31. **Fresnel Zone Method 263**
Huygens-Fresnel principle. Fresnel zones. Graphic computation of amplitude. Poisson's spot. Diffraction at the knife-edge of a semi-infinite screen. Zone plate as a lens. Drawbacks of the Fresnel zone method.

Sec. 32. **Kirchhoff's Approximation 269**
Green's formula. Helmholtz-Kirchhoff theorem. Radiation condition. Kirchhoff's approximation. Optical approximation. Fresnel-Kirchhoff diffraction relation. Helmholtz' reciprocity theorem. Secondary sources. Fresnel's approximation.

Sec. 33. **Fraunhofer Diffraction 277**
Fraunhofer diffraction region. Diffraction at a rectangular aperture. Diffraction at a slit. Diffraction at a circular aperture. Diffraction grating. Diffraction of white light at a grating. Dispersion region. Resolving power. Reflection gratings. Diffraction at a slit for a continuously varying wave phase. Phase gratings. Amplitude- and phase gratings. Oblique incidence of rays on a grating. Diffraction at continuous periodic and aperiodic structures. Diffraction by ultrasonic waves. Comparison of the characteristics of spectral instruments.

Sec. 34. **Fresnel Diffraction 294**
Fresnel's diffraction region. Diffraction at a rectangular aperture. Fresnel integrals. Cornu's spiral.

7. Basic Concepts of Fourier Optics

Sec. 35. **Lens as an Element Accomplishing a Fourier Transformation 299**
Phase transformation by a thin lens. Evaluation of the thickness function. Types of lenses. Lens as an element accomplishing a Fourier transformation.

Sec. 36. **Diffraction Formation of Images by a Lens 303**
Fourier transformation of amplitudes between the focal planes of a lens. Image formation by a lens. Limiting resolving power of optical instruments. Dark-field illumination method. Phase-contrast method.

Sec. 37. **Spatial Filtering of Images 313**
The essence of spatial filtering of images. Spatial filtering of the diffraction grating images. Abbe-Porter experiment.

Sec. 38. **Holography 315**
Synchronous detection. Hologram of a plane wave. Reconstruction of the image. Hologram of a point object. Hologram of an arbitrary object. Quality of a photographic plate and the exposure time. Three-dimensional reproduction of the object. Thick holograms (Denisyuk's method). Bragg's law. Recording of holograms and reconstruction of a plane wave. Recording of holograms and reconstruction of a spherical wave. Holographic recording and reconstruction of the image of an arbitrary object. Three-dimensional coloured image. Peculiarities of holograms as carriers of information. Applications of holography.

8. Propagation of Light in Anisotropic Media

Sec. 39. **Anisotropic Media 331**
Sources of anisotropy. Anisotropic dielectric media. Permittivity tensor.

Sec. 40. **Propagation of a Plane Electromagnetic Wave in an Anisotropic Medium 334**
Plane electromagnetic wave in anisotropic medium. Dependence of phase velocity on the direction of wave propagation and oscillations of vector D. Fresnel equation. Possible types of waves.

Sec. 41. **Passage of Rays Through an Anisotropic Medium 338**
Dependence of ray (group) velocity on direction. Ellipsoid of group velocities. Analysis of ray path through ellipsoid of group velocities. Optical axis. Biaxial and uniaxial crystals. Index ellipsoid. Ray surface.

Sec. 42. **Birefringence 344**
Ordinary and extraordinary rays. Essence of birefringence. Huygens' construction. Optical axis is perpendicular to the crystal surface. Optical axis is parallel to the crystal surface. Optical axis is at an angle to the crystal surface. Malus' law of rays. Polarization in birefringence. Polaroid. Polarizing prisms and birefringent prisms. Nicol prism. Birefringent prism. Polychroism.

Sec. 43. **Interference of Polarized Waves 349**
Interference of rays with mutually perpendicular directions of linear polarization. Quarter-wave plate. Half-wave plate. Wave plate. Analysis of linearly polarized light. Analysis of elliptically polarized light. Analysis of circularly polarized light. Compensators. Colours of crystal plates. Phenomena occurring in convergent rays.

Sec. 44. **Rotation of the Polarization Plane 354**
Rotation of polarization plane in crystalline bodies. Rotation of polarization plane in amorphous substances. Phenomenological theory of the rotation of polarization plane. Optical isomerism. Rotation of polarization plane in a magnetic field.

Sec. 45. **Artificial Anisotropy 358**
Anisotropy due to deformation. Anisotropy caused by an electric field. Anisotropy caused by a magnetic field. Pockels effect.

9. Scattering of Light

Sec. 46. **Scattering Processes** **365**
Nature of scattering. Types of scattering. Multiple scattering.

Sec. 47. **Rayleigh Scattering and Mie Scattering** **366**
Model of an elementary scatterer. Rayleigh scattering. Rayleigh law. Angular distribution and polarization of light in Rayleigh scattering. Attenuation of light intensity. Mie scattering. Angular intensity distribution and polarization of radiation in Mie scattering. Manifestations of Mie scattering.

Sec. 48. **Brillouin Scattering** **375**
Brillouin components. Undisplaced component. Brillouin scattering in solids.

Sec. 49. **Raman Scattering** **376**
Classical interpretation. Experimental facts. Quantum interpretation. Applications of Raman scattering.

10. Generation of Light

Sec. 50. **Blackbody Radiation** **381**
Radiation density. Equilibrium density of radiation. Kirchhoff's first law of radiation. Absorptive power and radiant emittance. Kirchhoff's second law of radiation. Blackbody. Number density of oscillation modes. Rayleigh-Jeans formula. Wien's formula. Planck's formula. Stefan-Boltzmann law. Wien's displacement law. Elementary quantum theory. Spontaneous and induced transitions. Einstein's coefficients.

Sec. 51. **Optical Amplifiers** **391**
Passage of light through a medium. Burger's law. Amplification conditions. Effect of light flux on the population density of levels. Saturation conditions. Creation of population inversion.

Sec. 52. **Lasers** **394**
Schematic diagram of a laser. Lasing threshold. Steady-state lasing conditions. Q-factor. Continuous-wave (CW) and pulsed lasers. Enhancement of the emission power. Q-switching method.

Sec. 53. **Laser Radiation** **398**
Radiation modes. Resonator with plane rectangular mirrors. Axial (longitudinal) modes. Width of emission lines. Side modes. Cylindrical resonator with spherical mirrors. Mode synchronization. Pulse duration. Attainment of mode synchronization. Laser speckles.

Sec. 54. **Characteristics of Some Lasers** **406**
Various types of lasers. Ruby laser. Helium-neon laser. Closed-volume CO_2-laser. CW-mode CO_2-laser. T-laser. Gasdynamic lasers. Dye lasers.

11. Nonlinear Phenomena in Optics

Sec. 55. **Nonlinear Polarization** **413**
Linear polarization. Nonlinear polarization. Quadratic nonlinearity. Nonlinear susceptibility. Combination frequencies.

Sec. 56. **Generation of Harmonics** **417**
Linear polarization waves. Nonlinear polarization waves. Spatial synchronism condition. Coherence length. Attainment of spatial synchronism. Generation of higher harmonics. Vector condition for spatial synchronism. Generation of sum and difference frequencies. Spontaneous decay of a photon. Parametric amplification of light. Parametric generators of light.

Sec. 57. Self-focussing of Light in a Nonlinear Medium 427
Nonlinear correction to the refractive index. Self-focussing and defocussing of a beam. Self-focussing length. Threshold energy flux. Main reasons behind nonlinearity of the refractive index. Time lag.

Appendix. SI units used in the book 432
Conclusion 434
Subject Index 441

Chapter 1

Electromagnetic Waves

Basic idea:
On the basis of the electromagnetic nature of light, the properties of monochromatic light waves are investigated with the help of Maxwell's equations.

SEC. 1. OPTICAL RANGE OF ELECTROMAGNETIC WAVES

The factors that make the waves of the visible range most suitable for vision are analyzed and the properties of the optical range are discussed.

WAVELENGTHS OF THE VISIBLE RANGE. The visible range includes electromagnetic waves that can be perceived by the human eye. The limiting range of these waves depends on the individual properties of the eye and varies approximately in the interval

$$\lambda = 0.38\text{-}0.76\ \mu\text{m}. \tag{1.1}$$

WAVE FREQUENCIES IN THE VISIBLE RANGE. In optics, one encounters the circular (cyclic) frequency

$$\omega = 2\pi/T, \tag{1.2}$$

where T is the oscillation period of the wave, as well as the frequency

$$\nu = 1/T. \tag{1.3}$$

These frequencies are connected through the obvious relation

$$\omega = 2\pi\nu. \tag{1.4}$$

The frequency is expressed in hertz, while the cyclic frequency is expressed in reciprocal seconds. Considering that

$$\lambda = cT, \tag{1.5}$$

where $c = 3 \times 10^8$ m/s is the velocity of light in vacuum, we obtain the following limiting values for the visible range:

$$\nu = (4\text{-}8) \times 10^{14}\ \text{Hz}, \tag{1.6}$$

$$\omega = (2.5\text{-}5.0) \times 10^{15}\ \text{s}^{-1}. \tag{1.7}$$

OPTICAL AND OTHER RANGES OF ELECTROMAGNETIC WAVES. It is possible to imagine theoretically that all frequencies from $\nu = 0$ to $\nu = \infty$ do exist. However, certain restrictions are imposed on these frequencies by the corpuscular properties of radiation. It is shown in the quantum theory that the electromagnetic radiation exists in the form of "packets" of energy (quanta). The energy of a quantum of radiation is connected with its frequency through the relation

$$\boxed{E = \hbar\omega = h\nu,} \tag{1.8}$$

where $h = 6.62 \times 10^{-34}$ J·s is called Planck's constant. The quantity $\hbar = h/(2\pi) = 1.05 \times 10^{-34}$ J·s is also called Planck's constant (new).

> Both these quantities are encountered equally often in physics according to circumstances. It is obvious from (1.8) that the constant $\hbar$ is convenient for use with the cyclic frequencies, while the constant h should be used for frequencies ν.

It is clear from (1.8) that infinite frequencies $\nu = \infty$ are not possible, since the corresponding radiation quanta would have infinite energies. The same relation also imposes a restriction on the lower values of the frequency if there exists the smallest possible value E_0 for the energy of a quantum. This means that the frequency cannot be lower than $\nu_0 = E_0/h$. At present, there are no indications in physics that the energy of the photons of electromagnetic radiation is restricted from below. Consequently, there is no lower limit to the frequency of the electromagnetic waves. The lowest frequency (about 8 Hz) has been observed for the standing electromagnetic waves between the ionosphere and the Earth's surface. This leads to the conclusion [see (1.8)] that the lowest energy of the electromagnetic radiation quanta is in any case below 10^{-33} J.

All possible frequencies of electromagnetic waves are subdivided into the following ranges (see the table below).

Each range has its own characteristic properties. As the wave frequency increases, the corpuscular properties of the radiation become more pronounced. Waves of different ranges

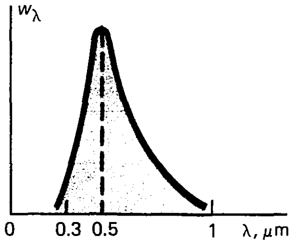


Fig. 1. Solar radiation spectrum.

Range	Bandwidth	
	in terms of wavelength λ	in terms of energy E of the quanta
Gamma radiation	< 0.0012 nm	> 1 MeV
X-rays	0.0012-12 nm	100 eV-1 MeV
Ultraviolet radiation	12-380 nm	3.2-100 eV
Visible radiation	380-760 nm	1.6-3.2 eV
Infrared radiation	760-10^6 nm = 1 mm	1.2×10^{-3}-1.6 eV
Radiowaves	> 1 mm	$< 1.2 \times 10^{-3}$ eV

can be also distinguished by the methods of generation of radiation. Each range forms the subject of investigation of the appropriate branch in physics.

The visible range and the adjoining ultraviolet and infrared regions together form the electromagnetic wave range which is studied in optics. The radiation quanta in the visible range are called photons and have an energy in the interval

$$E = (2.6\text{-}5.2) \times 10^{-19}\ \text{J} = 1.6\text{-}3.2\ \text{eV}. \tag{1.9}$$

WHY CAN WE SEE ONLY IN THE VISIBLE RANGE? At all temperatures above 0 K, all material bodies emit electromagnetic waves which are absorbed and reflected (scattered) by material objects. The intensity of the reflected and absorbed radiation depends on the radiation frequency, temperature, properties of the material, and other factors. The most intense source of electromagnetic radiation, which also determines the "radiation situation" on the Earth, is the Sun. The temperature of the solar surface is about 6000 K. Its radiation spectrum is the emission spectrum of a blackbody (see Sec. 50). The peak of the radiation spectrum on the scale of wavelengths lies at about 0.5 μm (Fig. 1).

Due to scattering and absorption while passing through the Earth's atmosphere, the solar spectrum is considerably altered depending on the thickness of the atmospheric layer, dust content, and other factors. Consequently, the spectrum at the Earth's surface terminates at a wavelength λ of about 0.3 μm. Waves with a smaller wavelength emitted by the Sun do not reach the surface of the Earth. This is due to their absorption by ozone (O_3) in the upper strata of the atmosphere. The wave attenuation mainly depends on the height of the Sun above the horizon. This can be graphically represented as follows:

!

The wavelengths of the visible range vary over the interval between 0.38 and 0.76 μm approximately, while the optical range also includes the infrared and ultraviolet spectral regions.
The visible range is most suitable for vision because an extremely low fraction of energy corresponds to the lower wavelengths during day time near the Earth's surface, while at higher wavelengths vision is hampered by noises.

?

What type of variations occur in the solar spectrum during the passage of solar radiation through the atmosphere? Indicate the boundaries of the visible range on the frequency scale as well as the circular frequency scale. What are the boundaries of the optical range of wavelengths?

Spectral range of light, Colour	Radiation fraction reaching the Earth for the position of the Sun		Wavelength, μm	Radiation energy, %	
	at the zenith	at dusk or dawn		in solar spectrum	near the Earth's surface
Red, 0.65 μm	0.96	0.21	0.3-0.4	5	1
Green, 0.52 μm	0.9	0.024	0.4-0.75	52	40
Violet, 0.41 μm	0.76	0.000065	0.75-2.3	43	59

Thus, the shortwave component of the spectrum is attenuated most strongly upon passing through the atmosphere. This is the reason behind the crimson hue of the Sun at dusk and dawn. On account of absorption in the atmosphere, the energy fraction of the ultraviolet part of the spectrum decreases while that of the infrared and microwave parts increases. For example, if the energy in the wavelength interval between 0.3 and 2.3 μm is taken as 100%, the energy distribution in the solar spectrum before penetrating the Earth's atmosphere and at the Earth's surface is characterized by the data given in the above table.

The visibility of the objects is due to the reflection of light. Hence the most suitable wavelength range for vision lies around the wavelength at which the radiation intensity has its peak, i. e. around $\lambda \simeq 0.5$ μm. Moreover, the interval should cover a considerable part of the total radiation energy. These conditions are fully satisfied by the visible range of the electromagnetic waves. It is quite natural that the faculty of the human eye to see was developed just in this range during the process of evolution.

However, the microwave part of the spectrum accounts for more than half the energy. Hence the possibility of human ability to see in this range cannot be ruled out in principle by energy considerations. Nevertheless, this spectral region is unsuitable for vision.

WHY IS MICROWAVE RANGE UNSUITABLE FOR VISION? The wave properties of radiation lead to the deterioration of the quality of vision with increasing wavelength because in this case the resolving power of optical instruments (see Sec. 36) as well as of the human eye deteriorate. Moreover, an increase in the wavelength necessitates an increase in the geometrical dimensions of receiving instruments, including the biological elements associated with vision. Naturally, this is a negative factor in the selection procedure during evolution.

However, the main factor which makes the microwave range unsuitable for vision is associated with the corpuscular nature of electromagnetic radiation and the existence of strong "noises" in this range, rendering vision impossible by the eye having a temperature of the order of body temperature of human beings or animals.

To begin with, we shall prove the impossibility of vision in the microwave range by considering the reflected solar radiation and then analyze the conditions under which it is possible to see objects in the microwave region ("night vision").

It is well known from molecular physics that the average number density of photons per oscillation mode (see Sec. 50) with a frequency ω in equilibrium radiation at a temperature T is given by

$$\langle n \rangle = \frac{1}{\exp [\hbar\omega/(kT)] - 1}, \tag{1.10}$$

where $k = 1.38 \times 10^{-23}$ J/K is the Boltzmann constant.

We are speaking of different types of oscillations having the same frequency but differing in their polarization propagation, direction of corresponding waves and other properties, as well as the oscillations of different frequencies. The average number density of photons having a frequency ω is equal to the sum of the number densities of photons, corresponding to different modes or types of oscillations at this frequency. In other words, this quantity is given by formula (1.10). A comparison of number densities of photons having various frequencies at different temperatures under equilibrium conditions boils down to a comparison of the average number densities (1.10), since the proportionality factor taking into account the number of modes is the same in all these cases.

Near the surface of the Earth, there are photons of solar radiation and photons emitted by the Earth's surface and by terrestrial objects. For the sake of definiteness, we assume that the temperature of the Earth's surface is $T_E = 300$ K. Then the number density of these photons per oscillation mode of frequency ω can be written as

$$\langle n_E \rangle = \frac{1}{\exp [\hbar\omega/(kT_E)] - 1}. \tag{1.11}$$

The temperature T of the surface of the Sun is about 6000 K. Near this surface, the number density of the equilibrium thermal radiation photons is given by formula (1.10) for $T = T_S$. The energy of radiation reaching the boundaries of the Earth's atmosphere has the same frequency distribution as near the surface of the Sun, but has a lower density. The spectral composition of the light passing through the Earth's atmosphere changes due to scattering and absorption. After this, light is reflected and absorbed by the surface of the Earth and terrestrial objects. Hence the spectral composition of the solar radiation has a very complex dependence on the conditions of its formation. To obtain an order-of-magnitude estimate of the spectral composition, however, all these variations can be neglected and it is assumed that the spectral composition of the radiation at the Earth's surface is approximately the same as of the original solar radiation. Hence the average number density $\langle n_S \rangle$ of the solar-radiation photons near the Earth's surface can be written in accordance with (1.10) as follows:

$$\langle n_S \rangle = \left(\frac{r_S}{R}\right)^2 \frac{1}{\exp\left[\hbar\omega/(kT_S)\right] - 1}, \tag{1.12}$$

where r_S is the radius of the Sun and R is the radius of the Earth's orbit. The factor $(r_S/R)^2$ takes into account the attenuation of the solar radiation density with increasing distance from the Sun. This attenuation is inversely proportional to the square of the distance. Since $r_S = 695.5 \times 10^3$ km and $R = 149.5 \times 10^6$ km, we obtain $(r_S/R)^2 \simeq 2.16 \times 10^{-5}$.

The decisive role for effective vision is played by the ratio of the flux of photons carrying information about an object to the flux of thermal photons which do not carry any information and merely create the background noise. The more the preeminence of the flux of information-carrying photons over the noise-generating photons, the better the conditions for vision. The flux of the photons carrying information about an object is formed by the scattering of the radiation of solar origin by the object. It can be assumed that this flux is proportional to the average number density $\langle n_S \rangle$ of photons. The flux of thermal photons that cause the background noise (we shall call them noise photons) is proportional to $\langle n_E \rangle$.

Let us estimate the efficiency of vision in the visible range. As an example, we consider photons having a wavelength $\lambda = 0.5\ \mu$m. The energy of these photons is $\hbar\omega = 2\pi\hbar c/\lambda \simeq 4 \times 10^{-19}$ J $\simeq 2.5$ eV. Considering that $kT_S = 8.3 \times 10^{-20}$ J $\simeq 0.52$ eV and $kT_E = 0.4 \times 10^{-20}$ J $\simeq 0.026$ eV, we obtain $\hbar\omega/(kT_S) = 4.8$ and $\hbar\omega/(kT_E) = 96$ for such photons.

According to (1.11) and (1.12), we obtain for the number density of photons

$$\langle n_E \rangle \simeq e^{-96}\ \mathrm{m}^{-3} \simeq 10^{-42}\ \mathrm{m}^{-3}, \tag{1.13}$$

$$\langle n_S \rangle \simeq 2 \times 10^{-5}\ e^{-4}\mathrm{m}^{-3} \simeq 10^{-7}\ \mathrm{m}^{-3}. \tag{1.14}$$

The ratio of the number of noise photons to the number of information-carrying photons is $\langle n_E \rangle / \langle n_S \rangle \simeq 10^{-35}$, which is a negligibly small quantity. There is practically no noise at this wavelength.

Let us now consider the situation in the microwave range (for example, at $\lambda = 2\ \mu m$). In this case, $\hbar\omega = 0.625$ eV, $\hbar\omega/(kT_S) \simeq 1.2$ and $\hbar\omega/(kT_E) \simeq 24$. Consequently,

$$\langle n_E \rangle \simeq e^{-24}\ m^{-3} \simeq 10^{-10}\ m^{-3}, \tag{1.15}$$

$$\langle n_S \rangle \simeq 2 \times 10^{-5} \times 0.43\ m^{-3} \simeq 10^{-5}\ m^{-3}. \tag{1.16}$$

Hence $\langle n_E \rangle / \langle n_S \rangle \simeq 10^{-5}$ (cf. the result mentioned above), i.e. the conditions for vision are much worse. However, it appears at first glance that in the absolute sense the situation is not so bad since in this mode each noise photon has 10^5 solar-radiation photons corresponding to it. Actually, this is not true. It must be borne in mind that vision is caused by rays reflected from the object. The rays from a point on the object reach the eye, forming a very small solid angle $\Delta\Omega \approx S/r^2$, where S is the area of the pupil and r is the distance between the eye and the object. Moreover, in this solid angle the reflected ray does not contain all possible modes of a given frequency, but only a few of them determined by the nature of reflection. Thermal photons, on the other hand, are present in the eye in all modes and fall on all points of the retina from all directions, i.e. from a solid angle of 2π. This increases the relative number of thermal photons by approximately a factor $\alpha = 2\pi/\Delta\Omega = 2\pi r^2/S$. Assuming that the pupil has a radius of 2 mm, we obtain $S = \pi (0.2)^2\ cm^2$. Hence for $r = 100$ cm, we get $\alpha = 0.5 \times 10^6$. Thus we have $\langle n_E \rangle / \langle n_S \rangle = 10$ instead of 10^{-5}, and this makes vision impossible at such wavelengths. All that has been said above about the relative increase in the number of noise quanta is naturally applicable to the visible range as well, but the extremely small value of the ratio $\langle n_E \rangle / \langle n_S \rangle \simeq 10^{-35}$ does not increase to an extent that would make vision impossible.

Summing up, we can state that the visible range is most suitable for vision since the shorter waves have too small energy while in the case of longer waves vision is hampered by thermal noises.

NIGHT VISION. At night, there are practically no solar-radiation photons near the Earth's surface (the possibility of lunar illumination and the glow of the sky at night are not taken into consideration). However, thermal radiation from material objects at a temperature of 300 K have their highest intensity around the wavelength $\lambda \simeq 10\ \mu m$. As it gets dark, thermodynamic equilibrium between the radiation and material bodies cannot be established since the conditions keep on changing continuously near the Earth's surface while the system is open from the side of the sky. Hence all the objects and the Earth's surface seem to "glow" at a wavelength $\lambda = 10\ \mu m$. The distribution of the emission energy over the spectrum depends to a considerable extent on the absorbing properties of the air and can be determined.

Let us suppose that the crystalline lens of the human eye can focus radiation with wavelength $\lambda = 10\ \mu m$ on the retina which is able to perceive this radiation and produce appropriate nerve pulses to cause the sense of vision. It can be asked whether a human being will be able to see the surrounding objects under such conditions. It follows from the above description that the flux of noise photons on any part of the retina considerably exceeds the flux of photons which form the image of an object on the retina. Hence the human eye will not be able to see the surrounding objects. To make vision possible, the density of noise photons in the human eye must be reduced, i. e. the temperature of the eye must be lowered. It has been shown by calculations that the temperatures required for this purpose are quite low and amount to just a few tens of kelvins. Hence night vision is possible only with the help of instruments maintained at a very low temperature. The image formed in the microwave region in these instruments is transformed into an image in the visible wavelength range and is observed by the eye at room temperature. During the image transformation, signals can be amplified or subjected to an appropriate treatment in order to obtain a high-quality visible image.

SEC. 2. PROPERTIES OF ELECTROMAGNETIC WAVES

The Maxwell equations are used to derive the basic properties of electromagnetic waves.

ELECTROMAGNETIC NATURE OF LIGHT. The existence of electromagnetic waves was theoretically predicted in 1862 by Maxwell (1831-1879) as a direct consequence of the electromagnetic field equations. The velocity of electromagnetic waves in vacuum was found to be equal (in present-day notation) to the quantity $1/\sqrt{\varepsilon_0\mu_0}$ which was termed the electrodynamic constant. The numerical value of this constant (3.1×10^8 m/s) was obtained earlier (in 1856) in the electromagnetic measurements of Weber (1804-1891) and Kohlrausch (1809-1858). This value was found to practically coincide with the velocity of light in vacuum which was measured in 1849 by Fizeau (1819-1896): $c = 3.15 \times 10^8$ m/s. Another important coincidence between the properties of electromagnetic waves and light is associated with the transverse nature of waves. Maxwell's equations show that electromagnetic waves are transverse waves, while the transverse nature of light waves follows from the experiments on polarization of light (Young, 1817). These two facts led Maxwell to the conclusion that light is a form of electromagnetic radiation.

The existence of electromagnetic waves was experimentally proved by Hertz (1857-1894) in 1888. The wavelength of the generated and detected electromagnetic waves was found to be about 66 cm. With the help of a metallic mirror, Hertz was able to observe the reflection and refraction of these waves, to study their polarization and to obtain standing waves, thus proving their ability to interfere.

WAVE EQUATION. In the absence of currents ($\mathrm{j} = 0$) and charges ($\rho = 0$), Maxwell's equations for vacuum have the following form:

$$\boxed{\operatorname{curl} \mathbf{H} = \partial \mathbf{D}/\partial t,} \tag{2.1}$$

$$\boxed{\operatorname{curl} \mathbf{E} = -\,\partial \mathbf{B}/\partial t,} \tag{2.2}$$

$$\boxed{\operatorname{div} \mathbf{B} = 0,} \tag{2.3}$$

$$\boxed{\operatorname{div} \mathbf{D} = 0,} \tag{2.4}$$

$$\boxed{\mathbf{D} = \varepsilon_0 \mathbf{E}, \ \mathbf{B} = \mu_0 \mathbf{H},} \tag{2.5}$$

where ε_0 and μ_0 are the electric and the magnetic constant respectively. Applying the curl operation to both sides of Eq. (2.1), we obtain

$$\frac{1}{\mu_0} \operatorname{curl} \operatorname{curl} \mathbf{B} = -\,\varepsilon_0 \frac{\partial}{\partial t} (\operatorname{curl} \mathbf{E}), \tag{2.6}$$

where we have taken into account Eq. (2.5) and the fact that the order of differentiation with respect to independent variables (spatial coordinates and time) can be changed. Considering the vector equality

$$\operatorname{curl} \operatorname{curl} \mathbf{B} = \operatorname{grad} \operatorname{div} \mathbf{B} - \nabla^2 \mathbf{B} \tag{2.7}$$

and replacing curl E on the right-hand side of (2.6) by its expression (2.2), we obtain the following equation for B:

$$\nabla^2 \mathbf{B} - \varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{B}}{\partial t^2} = 0. \tag{2.8}$$

Similarly, we obtain the following equation for E by applying the curl operation to both sides of Eq. (2.2):

$$\nabla^2 \mathbf{E} - \varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \tag{2.9}$$

The operator

$$\Box = \nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2}, \tag{2.10}$$

where $c = 1/\sqrt{\varepsilon_0 \mu_0}$ is the velocity of light in vacuum, is called the D'Alembert operator. With the help of this operator, we can write the wave equations (2.8) and (2.9) in the form:

$$\Box \mathbf{E} = 0, \quad \Box \mathbf{B} = 0. \tag{2.11}$$

PLANE WAVES. The wave equation plays a very significant role in physics. For a scalar function Φ, this equation has the form

$$\nabla^2 \Phi - \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2} = 0. \tag{2.12}$$

Let us find the general solution of this equation for the case when Φ depends on only one of the Cartesian coordinates, say, z, i.e. $\Phi = \Phi(z, t)$. This means that Φ has a constant value at all points on a plane perpendicular to the Z-axis. In this case, Eq. (2.12) assumes the form

$$\frac{\partial^2 \Phi}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2} = 0. \tag{2.13}$$

Using new independent variables

$$\xi = z - ct, \quad \eta = z + ct, \tag{2.14}$$

we obtain

$$\frac{\partial \Phi}{\partial z} = \frac{\partial \Phi}{\partial \xi} \frac{\partial \xi}{\partial z} + \frac{\partial \Phi}{\partial \eta} \frac{\partial \eta}{\partial z} = \frac{\partial \Phi}{\partial \xi} + \frac{\partial \Phi}{\partial \eta}, \tag{2.15a}$$

$$\frac{\partial \Phi}{\partial t} = \frac{\partial \Phi}{\partial \xi} \frac{\partial \xi}{\partial t} + \frac{\partial \Phi}{\partial \eta} \frac{\partial \eta}{\partial t} = -c \frac{\partial \Phi}{\partial \xi} + c \frac{\partial \Phi}{\partial \eta}. \tag{2.15b}$$

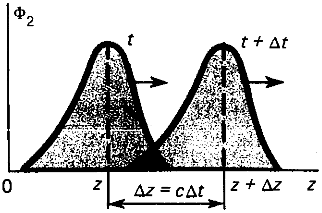


Fig. 2. A wave moving in the positive direction of the Z-axis: $\Phi = \Phi_2(z - ct)$.

Dividing both sides of Eq. (2.15b) by c and subtracting them termwise from both sides of Eq. (2.15a), we obtain

$$\frac{\partial}{\partial z} - \frac{1}{c}\frac{\partial}{\partial t} = 2\frac{\partial}{\partial \xi}. \tag{2.16}$$

Similarly, the termwise addition of both sides of the same two equations gives

$$\frac{\partial}{\partial z} + \frac{1}{c}\frac{\partial}{\partial t} = 2\frac{\partial}{\partial \eta}. \tag{2.17}$$

Thus,

$$\left(\frac{\partial}{\partial z} - \frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial z} + \frac{1}{c}\frac{\partial}{\partial t}\right) = \frac{\partial^2}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}.$$

Using (2.16) and (2.17), we can transform Eq. (2.13) as follows:

$$\frac{\partial^2\Phi}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 4\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta}\Phi = 0. \tag{2.18}$$

Integrating (2.18) with respect to ξ, we obtain a function independent of ξ, which can depend in the present case only on η, i.e. is an arbitrary function $\Psi(\eta)$. Then Eq. (2.18) assumes the form

$$\frac{\partial\Phi}{\partial\eta} = \Psi(\eta). \tag{2.19}$$

Integrating this equation with respect to η, we obtain

$$\Phi = \int \Psi(\eta)\,\mathrm{d}\eta = \Phi_1(\eta) + \Phi_2(\xi), \tag{2.20}$$

where $\Phi_1(\eta)$ is the primitive in the integral of $\Psi(\eta)$ with respect to $\mathrm{d}\eta$ and $\Phi_2(\xi)$ is the integration constant. It can be seen from the nature of the solution that the functions Φ_1 and Φ_2 are arbitrary. Taking (2.14) into account, we can write the solution (2.20) of Eq. (2.13) in the form

$$\Phi(z,t) = \Phi_1(z + ct) + \Phi_2(z - ct). \tag{2.21}$$

Let us find out the physical meaning of this solution. First we analyze the solution

$$\Phi = \Phi_2(z - ct). \tag{2.22}$$

Figure 2 shows a plot of Φ_2 as a function of z at the instants of time t and $t + \Delta t$. It can be seen that the argument of the function Φ_2 at a point z at instant t coincides with the

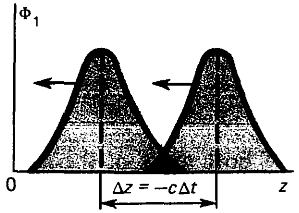


Fig. 3. A wave moving in the negative direction of the Z-axis: $\Phi = \Phi_1(z + ct)$.

argument of this function at the point $z + \Delta z$ at the instant $t + \Delta t$ if $\Delta z = c\Delta t$, since

$$z - ct = z + \Delta z - c(t + \Delta t) \quad (\Delta z = c\Delta t). \tag{2.23}$$

Hence the plot of the function for $t + \Delta t$ is obtained from the plot for t by displacing all points of the curve in the positive direction of the Z-axis by $\Delta z = c\Delta t$. Hence the wave velocity is given by $v = \Delta z/\Delta t = c$. The function $\Phi_2\ (z - ct)$ describes a wave of arbitrary shape moving with a velocity c in the positive direction of the Z-axis. During propagation of the wave, the value of Φ_2 at each point and the shape of the wave remain unchanged.

The physical meaning of Φ_1, i.e. of the solution

$$\boxed{\Phi = \Phi_1(z + ct),} \tag{2.24}$$

is explained in the same way. Considering that

$$z + ct = z + \Delta z + c(t + \Delta t) \quad (\Delta z = -c\Delta t), \tag{2.25}$$

we conclude that the function $\Phi_1(z + ct)$ describes a wave of arbitrary shape, moving with a velocity c in the negative direction of the Z-axis (Fig. 3). The value of Φ_1 at each point of the wave and the shape of the wave do not change in the process of wave propagation.

The wave described by formula (2.21) is a superposition of two waves moving in opposite directions. In this case, we cannot speak of the velocity or direction of the wave. In the simplest case, we obtain a standing wave, while in the general case we get a complex electromagnetic field which requires a special investigation.

For fixed values of z and t, the values of the function Φ are constant in a plane perpendicular to the Z-axis. Hence such waves are called plane waves.

SPHERICAL WAVES. If the wave emanating from a point source is isotropic, the solution of Eq. (2.12) must be sought in the form $\Phi = \Phi(r, t)$, where r is the distance from the point source which is taken as the origin of coordinates. Considering that in the system of spherical coordinates (r, θ, φ)

$$\nabla^2\Phi = \frac{1}{r}\frac{\partial^2}{\partial r^2}(r\Phi) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial\theta} \times \left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial\varphi^2}, \tag{2.26}$$

and the required solution is independent of angular variables, Eq. (2.12) can be written in the form

?

Why cannot a universal relation be established between the frequency of a wave and the wave number?
Which universal relation connects the frequency and the wave number in an isotropic medium with the uniform velocity of wave propagation?
What causes the invariance of a plane wave?
How are the frequency and wave vector of a wave transformed as it passes from one inertial reference system to another?

$$\frac{\partial^2}{\partial r^2}(r\Phi) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}(r\Phi) = 0\,, \tag{2.27}$$

which is the same as (2.13) but for $z \rightarrow r$ and $\Phi \rightarrow r\Phi$. Hence we can write instead of (2.21)

$$r\Phi(r, t) = \Phi_1(r + ct) + \Phi_2(r - ct)\,,$$

where Φ_1 and Φ_2, as before, are arbitrary functions of their arguments. Consequently, the general spherically symmetric solution of Eq. (2.12) has the form

$$\boxed{\Phi(r, t) = \frac{\Phi_1(r + ct)}{r} + \frac{\Phi_2(r - ct)}{r}.} \tag{2.28}$$

The physical meaning of the individual components of (2.26) is explained in the same way as for (2.21). The second term represents a wave moving in the direction of increasing r, i.e. away from the centre. Such a wave is called a divergent wave. The first term describes a wave moving in the direction of decreasing r, i.e. towards the centre. Such a wave is called a convergent wave. The general solution (2.28) is a superposition of a divergent and a convergent wave.

For a sphere of constant radius, the value of Φ is constant at a fixed instant of time. Such waves are called spherical waves.

PLANE HARMONIC WAVES. If Φ_1 and Φ_2 in (2.21) are harmonic functions of their argument, the wave is called a harmonic wave. For example, we can write the function Φ_2 in the form

$$\Phi_2(z - ct) = \Phi_2[-c(t - z/c)] = A\cos\omega(t - z/c)\,,$$

where A is a constant and ω is the frequency of the harmonic function. The wave described by the function

$$\boxed{\Phi(z, t) = A\cos\omega(t - z/c)\,,} \tag{2.29}$$

is called a plane harmonic wave. It propagates in the positive direction of the Z-axis. The constant A is called the wave amplitude, while ω is called the frequency of the wave. Since the wave is in motion, it is also called a travelling wave. An expression similar to (2.29) can be written for a wave by using the sine. The general expression for a travelling wave propagating in the positive direction of the Z-axis has the form

$$\boxed{\Phi(z, t) = A\cos\omega(t - z/c) + B\sin\omega(t - z/c).} \tag{2.30}$$

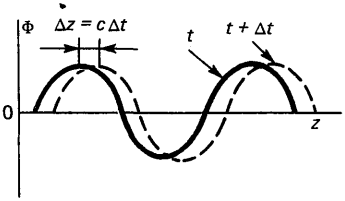


Fig. 4. A plane harmonic wave at two successive instants of time.

By a suitable choice of the initial instant of time, a travelling plane wave can always be represented in the form (2.29) or by an analogous expression with the sine. Obviously, the wave travelling in the negative direction of the Z-axis can be described by the function

$$\Phi(z,t) = A\cos\omega(t + z/c) + B\sin\omega(t + z/c). \tag{2.31}$$

Figure 4 shows a plane harmonic wave at two successive instants of time t and $t + \Delta t$. For the sake of clarity in representation, it can be assumed that this is a wave on the surface of water, and that Φ represents the deviation of the particles of the water surface from the horizontal plane. Of course, for such a representation c is not the velocity of light but the rate of wave propagation relative to water. Positive values of Φ correspond to "crests" on the water surface, while negative values correspond to "troughs". The figure shows just a small part of the wave containing two "crests" and one "trough". If we observe any fixed point of the medium, we find that it oscillates in time according to a harmonic law. For example, this law can be described by the function

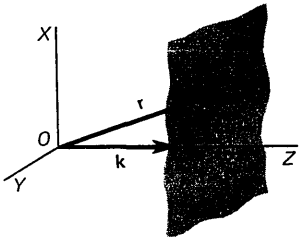


Fig. 5. To the vector notation of a plane wave.

$$\Phi(0,t) = A\cos\omega t. \tag{2.32}$$

If the observer "sits" at a certain point of the wave, say, at the top of a "crest", and moves along with the wave, he will not observe any changes in the wave pattern and will always see the same time-independent set of "crests" and "troughs" constituting the plane wave.

The argument of the harmonic function in (2.29) is called the phase of the wave. A wave whose constant-phase surfaces are planes is called a plane wave. Considering that

$$\lambda = cT = 2\pi c/\omega, \tag{2.33}$$

we can write (2.29) in the form

$$\Phi(z,t) = A\cos(\omega t - \omega z/c) = A\cos(\omega t - kz), \tag{2.34}$$

where $k = \omega/c = 2\pi/\lambda$ is the wave number.

WAVE VECTOR. In order to get rid of the system of coordinates, we can write (2.34) in vector notation. Let vector $\mathbf{k}$ have a magnitude equal to the wave number and a direction parallel to the positive direction of the Z-axis (Fig. 5). Such a vector is called a wave vector. Considering that $\mathbf{k}\cdot\mathbf{r} = kz$, we can write the following expression instead of (2.34) for an arbitrary point characterized by the radius vector $\mathbf{r}$:

$$\Phi(\mathbf{r},t) = A\cos(\omega t - \mathbf{k}\cdot\mathbf{r}). \tag{2.35}$$

This formula is independent of the system of coordinates and characterizes a plane wave propagating in the direction of vector k.

An expression for a wave similar to (2.35) can be also written in terms of the sine:

$$\Phi'(\mathbf{r}, t) = A' \sin(\omega t - \mathbf{k}\cdot\mathbf{r}). \tag{2.36}$$

Since $\sin(\alpha + \pi/2) = -\cos\alpha$, a suitable choice of the initial instant of time will again reduce this expression to (2.35).

COMPLEX REPRESENTATION OF A PLANE WAVE. Using Euler's formula

$$e^{i\alpha} = \cos\alpha + i \sin\alpha, \tag{2.37}$$

we can express (2.35) and (2.36) in the form

$$\Phi(\mathbf{r}, t) = A \operatorname{Re}[e^{-i(\omega t - \mathbf{k}\cdot\mathbf{r})}], \tag{2.38a}$$

$$\Phi'(\mathbf{r}, t) = -A' \operatorname{Im}[e^{-i(\omega t - \mathbf{k}\cdot\mathbf{r})}], \tag{2.38b}$$

where Re and Im stand for real and imaginary parts of a complex number. For calculations, it is convenient to use the complex representation of a plane wave in the form

$$\Phi(\mathbf{r}, t) = A e^{-i(\omega t - \mathbf{k}\cdot\mathbf{r})}, \tag{2.39}$$

denoting the complex and real quantities by the same symbol. This simplifies the formulas and does not cause any confusion. If, however, a confusion is likely to arise, we shall explicitly indicate as to which representation is meant.

The quantity A in (2.39) can be real, complex or imaginary. Considering that in the general case

$$A = |A| e^{i\varphi}, \quad \tan\varphi = \operatorname{Im} A / \operatorname{Re} A, \tag{2.40}$$

we can write (2.39) in the form

$$\Phi(\mathbf{r}, t) = |A| e^{-i(\omega t - \mathbf{k}\cdot\mathbf{r} - \varphi)}, \tag{2.41}$$

where $|A|$ is the amplitude of the plane wave. Hence in (2.39) also, $|A|$ is the amplitude of the plane wave and $\omega t - \mathbf{k}\cdot\mathbf{r} - \varphi$ is the phase, where $\tan\varphi = \operatorname{Im} A / \operatorname{Re} A$.

COMPLEX REPRESENTATION OF A SPHERICAL WAVE. From the method of recording a plane wave in the complex form, it is obvious that divergent and convergent spherical waves represented by (2.28) can be expressed in the following forms respectively:

$$\Phi(r,t) = (A/r)\,e^{-i(\omega t - kr)}, \tag{2.42}$$

$$\Phi(r,t) = (A/r)\,e^{-i(\omega t + kr)}. \tag{2.43}$$

It should be noted that the exponential indices in these formulas contain the product of positive quantities k and r rather than the scalar product $\mathbf{k}\cdot\mathbf{r}$.

PLANE ELECTROMAGNETIC WAVE. To analyze the structure of a plane electromagnetic wave, it is convenient to use the operator

$$\nabla = \mathbf{i}_x \frac{\partial}{\partial x} + \mathbf{i}_y \frac{\partial}{\partial y} + \mathbf{i}_z \frac{\partial}{\partial z} \tag{2.44}$$

to write Maxwell's equations in symbolic form. Considering that

$$\operatorname{curl} \mathbf{A} = \nabla \times \mathbf{A}, \qquad \operatorname{div} \mathbf{A} = \nabla \cdot \mathbf{A}, \tag{2.45}$$

we can write Maxwell's equations (2.1)-(2.5) in the following form:

$$\nabla \times \mathbf{B} = \mu_0 \varepsilon_0 \partial \mathbf{E}/\partial t, \tag{2.46}$$

$$\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t, \tag{2.47}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{2.48}$$

$$\nabla \cdot \mathbf{E} = 0. \tag{2.49}$$

We shall seek the solution of these equations in the form

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{-i(\omega t - \mathbf{k}\cdot\mathbf{r})}, \tag{2.50}$$

$$\mathbf{B}(\mathbf{r},t) = \mathbf{B}_0 e^{-i(\omega t - \mathbf{k}\cdot\mathbf{r})}, \tag{2.51}$$

where $\mathbf{E}_0$ and $\mathbf{B}_0$ are constant vectors which are independent of coordinates and time. These vectors may have complex components.

Substituting expressions (2.50) and (2.51) into Eqs. (2.46)-(2.49) and considering that

$$\nabla e^{i\mathbf{k}\cdot\mathbf{r}} = i\mathbf{k}e^{i\mathbf{k}\cdot\mathbf{r}}, \quad \frac{\partial}{\partial t} e^{-i\omega t} = -i\omega e^{-i\omega t}, \tag{2.52}$$

we obtain the following relations:

$$-\mathbf{k} \times \mathbf{B} = \omega \mu_0 \varepsilon_0 \mathbf{E}, \tag{2.53}$$

$$\mathbf{k} \times \mathbf{E} = \omega \mathbf{B}, \tag{2.54}$$

$$\mathrm{k}\cdot\mathrm{B}=0, \tag{2.55}$$

$$\mathrm{k}\cdot\mathrm{E}=0. \tag{2.56}$$

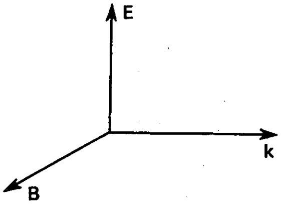


Fig. 6. Mutual orientation of the vectors of a plane electromagnetic wave.

It follows from relations (2.55) and (2.56) that the vectors E and B of a plane wave are perpendicular to the vector k, i. e. to the direction of wave propagation. This means that the electromagnetic wave is a transverse wave. Relations (2.53) and (2.54) show that the vectors E and B are mutually perpendicular. Thus, E, B and k form a set of three mutually perpendicular vectors (Fig. 6).

The transverse nature of light oscillations was discovered in 1817 by Young (1773-1829). He used this concept to explain the absence of interference between light rays polarized in mutually perpendicular planes, an effect which was experimentally discovered in 1816 jointly by Arago (1786-1853) and Fresnel (1788-1827).

Taking the modulus of both sides of (2.54), $|\mathrm{k}\times\mathrm{E}|=\omega|\mathrm{B}|$, and considering that $|\mathrm{k}\times\mathrm{E}|=|\mathrm{k}||\mathrm{E}|$, $|\mathrm{k}|=k=\omega/c$, we obtain the following relation between the electric field strength and the magnetic induction of a plane wave in vacuum:

$$\boxed{E=cB.} \tag{2.57}$$

Since k, ω, μ_0 and ε_0 in (2.53) and (2.54) are real quantities, we can conclude from (2.50) and (2.51) that E and B change in a plane wave in phase, i. e. they simultaneously attain their peak and zero values (Fig. 7).

INVARIANCE OF A PLANE WAVE. In accordance with the basic properties of a plane wave, the vectors E and B of the wave are mutually perpendicular and relation (2.57) between them is satisfied. First of all, the question of invariance of a plane wave arises, i. e. whether a wave which is plane in one system of coordinates is also plane in all other coordinate systems which move uniformly in a straight line relative to the first system. The answer to this question is based on the electromagnetic field invariants. It was shown in *Electricity and Magnetism** that in Lorentz transformations the following quantities characterizing an electromagnetic field are invariant

$$\boxed{I_1=c^2B^2-E^2, \qquad I_2=\mathrm{E}\cdot\mathrm{B}.} \tag{2.58}$$

If a plane wave describes an electromagnetic field in some system of coordinates, then $I_1=0$ and $I_2=0$. Since these quantities are invariant in Lorentz transformations, they are

---

* A. N. Matveev. *Electricity and Magnetism*, Mir Publishers, Moscow, 1986.

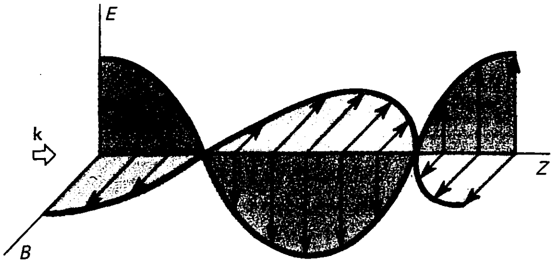


Fig. 7. Variation of the vectors of a plane wave in space.

also equal to zero in all other coordinate systems. According to (2.58) this means that the electromagnetic field is a plane wave in all other coordinate systems as well. This proves the invariance of a plane wave.

PHASE INVARIANCE. Obviously, the statement that the field vector of a wave is equal to zero at some point in space and time has a physical meaning irrespective of the system of coordinates in which this point is considered. In other words, the field vectors at this point in space and time are equal to zero in all systems of coordinates. This means that the phase of the wave is the same in all systems of coordinates. This proves the phase invariance, which follows from the formulas for field vector transformations. Having written the formulas for the transformation of the electromagnetic field vectors of a plane wave and substituting into them expressions of the type (2.50) and (2.51), we can conclude at once that phase invariance of the waves is essential for the validity of the transformation formulas at any point in space and time.

FOUR-DIMENSIONAL WAVE VECTOR. We can express the phase in the form

$$\mathbf{k}\cdot\mathbf{r} - \omega t = k_1 x_1 + k_2 x_2 + k_3 x_3 + k_4 x_4, \tag{2.59}$$

where

$$x_1 = x, \quad x_2 = y, \quad x_3 = z, \quad x_4 = ic, \tag{2.60a}$$

$$k_1 = k_x, \quad k_2 = k_y, \quad k_3 = k_z, \quad k_4 = i\omega/c. \tag{2.60b}$$

The right-hand side of (2.59) has the form of a scalar product. Since the set of quantities in (2.60a) constitutes a four-dimensional vector, the invariance of the scalar product on the right-hand side of (2.59) (phase invariance) means that the set of quantities in (2.60b) also constitutes a four-dimensional vector. Using the familiar formulas from the theory of relativity, we can transform these quantities from one system of coordinates into another.

TRANSFORMATION FORMULAS FOR FREQUENCY AND DIRECTION OF PROPAGATION OF A PLANE

WAVE. Denoting by n a unit vector in the direction of propagation of a plane wave, we can represent the wave vector in the form

$$\mathbf{k} = \mathbf{n}\omega/c.$$

Since the set (k_1, k_2, k_3, k_4) constitutes a four-dimensional vector, we obtain the following transformation formulas:

$$\begin{gathered}\omega n_x = \omega'(\beta + n_x')/\sqrt{1-\beta^2}, \quad \omega n_y = \omega' n_y', \\ \omega n_z = \omega' n_z', \quad \omega = \omega'(1 + \beta n_x')/\sqrt{1-\beta^2},\end{gathered} \tag{2.61}$$

where the primed quantities refer to the system of coordinates K' moving relative to the fixed coordinate system K in the positive direction of the X-axis with a velocity v ($\beta = v/c$). The X- and X'-axes coincide, while the axes Z' and Y' are parallel to the Z- and Y-axes respectively.

DOPPLER EFFECT. According to formulas (2.61), the frequency depends on the relative velocity of the source and the observer. Using the relativity principle and the last equation in (2.61), we obtain the following formula

$$\omega = \omega' \sqrt{1-\beta^2}/(1 - \beta n_x). \tag{2.62}$$

This formula describes the effect discovered in 1842 by Doppler (1803-1853) and named after him. At low velocities ($\beta = v/c \ll 1$), formula (2.62) assumes a simpler form:

$$\omega = \omega'(1 + \beta n_x). \tag{2.63}$$

If a wave propagates along the X-axis, $n_x = \pm 1$ and (2.63) assumes the form

$$\Delta\omega/\omega' = (\omega - \omega')/\omega' = \pm v/c. \tag{2.64}$$

The sign "+" corresponds to the case when the source approaches the observer, while "−" corresponds to the case when the source is receding from the observer. This mode of frequency variation is called the longitudinal Doppler effect.

If the source is moving in a direction perpendicular to the direction of motion of the observer ($n_x = 0$), the transverse Doppler effect is observed:

$$\omega = \omega' \sqrt{1-\beta^2}. \tag{2.65}$$

This effect is due to the factor $\sqrt{1-\beta^2}$ in (2.62) and is a purely relativistic effect associated with the time dilation effect for a moving observer. This effect was confirmed experimentally.

Example 2.1. A mirror moves with a velocity v in the positive direction of the X-axis so that its reflecting plane surface is perpendicular to the X-axis. A ray of light is incident on the mirror at an angle $\theta_0 < \pi/2$ to the positive direction of the X-axis. Find the direction of the reflected ray if it lies in the XY-plane.

The projections of the unit vector n directed along the ray incident on the mirror are equal to $(\cos\theta_0, \sin\theta_0, 0)$. If the mirror were fixed, the projections of the unit vector characterizing the reflected ray would be $(-\cos\theta_0, \sin\theta_0, 0)$ in view of the fact that the angle of reflection is equal to the angle of incidence. The effect becomes more complicated if the mirror is in motion.

Let us solve the problem by using the relativistic principle. We consider the reflection in the system K' in which the mirror is at rest and the law of reflection is known. From (2.61) and (2.62), we obtain the formulas for transformation from the system K to the system K':

$$\begin{aligned} \omega' &= \omega_0(1-\beta n_x)/\sqrt{1-\beta^2} = \omega_0(1-\beta\cos\theta_0)/\sqrt{1-\beta^2}, \\ n'_x &= (n_x-\beta)/(1-\beta n_x) = (\cos\theta_0-\beta)/(1-\beta\cos\theta_0), \qquad (2.66) \\ n'_y &= n_y\sqrt{1-\beta^2}/(1-\beta n_x) = \sin\theta_0\sqrt{1-\beta^2}/(1-\beta\cos\theta_0), \end{aligned}$$

where ω_0 is the frequency of light in the system K before reflection from the moving mirror. Upon reflection from the fixed mirror in the system K', the frequency does not change and the angle of reflection is equal to the angle of incidence. Hence we can write the following relations for the reflected ray on the basis of (2.66):

$$\begin{aligned} \omega'_{\text{refl}} &= \omega_0(1-\beta\cos\theta_0)/\sqrt{1-\beta^2}, \\ n'_{x\,\text{refl}} &= -(\cos\theta_0-\beta)/(1-\beta\cos\theta_0), \qquad (2.67) \\ n'_{y\,\text{refl}} &= \sin\theta_0\sqrt{1-\beta^2}/(1-\beta\cos\theta_0). \end{aligned}$$

We can now revert to the system K with the help of formulas (2.61):

$$\begin{aligned} \omega_{\text{refl}} &= \omega'_{\text{refl}}\frac{1+\beta n'_x}{\sqrt{1-\beta^2}} = \omega_0\frac{1-2\beta\cos\theta_0+\beta^2}{1-\beta^2}, \\ \cos\theta_{\text{refl}} &= n_{x\,\text{refl}} = \frac{\beta+n'_{x\,\text{refl}}}{1+\beta n'_{x\,\text{refl}}} \\ &= -\frac{\cos\theta_0-2\beta+\beta^2\cos\theta_0}{1-2\beta\cos\theta_0+\beta^2}, \qquad (2.68) \\ \sin\theta_{\text{refl}} &= n_{y\,\text{refl}} = \frac{n'_{y\,\text{refl}}\sqrt{1-\beta^2}}{1+\beta n'_{x\,\text{refl}}} \\ &= \frac{\sin\theta_0(1-\beta^2)}{1-2\beta\cos\theta_0+\beta^2}. \end{aligned}$$

!

The existence of electromagnetic waves is a consequence of Maxwell's equations. The electromagnetic nature of light was established from the coincidence of the properties of the electromagnetic waves described by Maxwell's equations, and the properties of light.

Waves are classified according to the shape of the constant phase surfaces (plane, spherical, cylindrical, etc.).

Harmonic waves are the simplest and most important among electromagnetic waves, although other types of waves also exist.

The complex representation of waves is very convenient for calculations.

In the nonrelativistic case, when $\beta \ll 1$, these equalities can be represented in the following form with an error not exceeding β in order of magnitude:

$$\omega_{\text{refl}} \approx \omega_0 (1 - 2\beta \cos\theta_0), \quad \cos\theta_{\text{refl}} \approx -\cos\theta_0 + 2\beta \sin^2\theta_0,$$
$$\sin\theta_{\text{refl}} \approx \sin\theta_0 (1 + 2\beta\cos\theta_0). \tag{2.69}$$

It can be seen that a ray is reflected from a moving mirror at a larger angle than from a fixed mirror.

SEC. 3. FLUX DENSITIES OF ENERGY AND MOMENTUM OF ELECTROMAGNETIC WAVES. LIGHT PRESSURE

General formulas for the flux density of the energy and momentum of electromagnetic field are elaborated for the optical range.

ENERGY FLUX DENSITY. The flux density of the energy of electromagnetic field is defined by Poynting's vector

$$\boxed{\mathbf{S} = \mathbf{E} \times \mathbf{H},} \tag{3.1}$$

whose magnitude for a plane wave can be presented in the form

$$S = |\mathbf{S}| = |\mathbf{E} \times \mathbf{H}| = |\mathbf{E}||\mathbf{H}| = EB/\mu_0 = [1/(\mu_0 c)]E^2, \tag{3.2}$$

where we have taken into account formula (2.57). Considering that $1/\mu_0 = \varepsilon_0 c^2$, we can write relation (3.2) in the form

$$\boxed{S = c\varepsilon_0 E^2.} \tag{3.3}$$

Formulas (3.1)-(3.3) contain instantaneous values of quantities. However, the electromagnetic wave vectors in the optical range oscillate with frequencies of the order of 10^{15} Hz, and hence it is impossible to observe the variation of quantities with time. We can observe and measure only the mean value of quantities over a very large number of periods of oscillations. Hence we must change over from instantaneous to mean values.

Considering that $E = E_0 \cos\omega t$, where E_0 is the amplitude of the electric field strength of the wave, we obtain the energy flux density averaged over time:

$$\boxed{\langle S\rangle_t = c\varepsilon_0 E_0^2 \langle\cos^2\omega t\rangle_t = c\cdot(1/2)\varepsilon_0 E_0^2} \tag{3.4}$$

(the subscript t at the angle brackets indicates that the averaging is carried out over time. In most cases, we shall omit this subscript since it is clear that we are always referring to time-averaging).

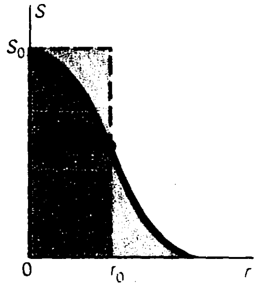


Fig. 8. Gaussian (solid curve) and effective (dashed line) distributions of the energy flux density in a beam.

FLUX DENSITY DISTRIBUTION OVER BEAM CROSS SECTION. In experiments, one frequently employs light beams of a finite cross section over which the flux density is distributed nonuniformly. By definition, the rate of flow of electromagnetic energy in a beam is given by

$$P = \int_{\sigma} |S| \mathrm{d}\sigma, \tag{3.5}$$

where σ is the cross-sectional area of the beam in a direction perpendicular to the direction of propagation of light. Since the beam cross section is usually of the order of 1 mm-1 cm, i.e. $r_0 \simeq 10^3$-10^4 μm, while the wavelength is of the order of 1 μm, we conclude that $\lambda/r_0 \simeq 10^{-3}$-$10^{-4}$. Consequently, the diffraction effects (see Sec. 33) in the beams are insignificant and the energy flux density can be calculated by using formulas for plane electromagnetic waves.

GAUSSIAN BEAM. Most often, a beam has a circular cross section over which the energy density distribution is axially symmetric and Gaussian. Such a beam is called a Gaussian beam, and the amplitudes of the plane electromagnetic waves constituting this beam are distributed according to the law

$$E_0(r) = A_0 \mathrm{e}^{-r^2/(2r_0^2)}, \tag{3.6}$$

where A_0 is a constant equal to the amplitude of the waves at the centre of the beam ($r = 0$), and r is the distance from the centre. The mean energy flux density is given by

$$S = (1/2)\varepsilon_0 A_0^2 \mathrm{e}^{-r^2/r_0^2} = S_0 \mathrm{e}^{-r^2/r_0^2}, \tag{3.7}$$

where $S_0 = \varepsilon_0 A_0^2/2$ is the mean energy flux density at the centre of the beam. To simplify the notation, the angle brackets indicating the mean value of the quantity have been omitted.

In accordance with (3.5), the rate of flow of energy in the beam is given by the expression

$$P = S_0 \int_0^{\infty} \mathrm{e}^{-r^2/r_0^2} \cdot 2\pi r \mathrm{d}r = S_0 \pi r_0^2. \tag{3.8}$$

Actually, the energy density in a Gaussian beam is distributed according to exponential law (3.7). The density decreases by a factor e = 2.7 at a distance r_0. By convention concerning the use of exponentially decreasing quantities, it can be stated that the beam radius is equal to r_0. In accordance with (3.8), it is assumed for rough estimates that the entire beam energy is concentrated in a circular cross section of radius r_0, while the density over the cross section is constant and equal to the density at the centre of the beam. The Gaussian distribution of the flux density and the effective distribution are shown in Fig. 8 by a solid curve and a dashed line respectively. Such an approach is usually extended to other (non-Gaussian) beams as well. The rate of flow of energy of a beam is expressed in the form

$$P = S_{\mathrm{eff}} \pi r_{\mathrm{eff}}^2, \tag{3.9}$$

where S_{eff} is the effective energy flux density in the beam and r_{eff} is the effective beam radius. If a light pulse has an energy W, its effective duration τ_{eff} is connected with the rate of flow of energy in the pulse through the relation

$$W = P\tau_{\mathrm{eff}}. \tag{3.10}$$

At present, enormous energy flux densities S of the order of 10^{20} W/m^2 have been obtained with the help of lasers. This means [see (3.3)] that the electric field strengths E in the wave are of the order of 10^9 W/m, i.e. assume the values typical for intraatomic fields. The duration τ of pulses is 10^{-12} s or less.

MOMENTUM DENSITY OF ELECTROMAGNETIC WAVES. An electromagnetic wave possesses energy as well as momentum. It was shown in the theory of electricity and magnetism that the momentum density $\mathbf{G}$ of an electromagnetic wave is connected with the energy flux density $\mathbf{S}$ in it through the relation

$$\boxed{\mathbf{G} = \mathbf{S}/c^2.} \tag{3.11}$$

LIGHT PRESSURE. When light is scattered or absorbed by a body, the latter acquires a certain momentum in accordance with the conservation law. This momentum is equal to the difference in the momenta of the light beam before and after scattering or absorption. As a result, the body is acted upon by a certain force and a light pressure is exerted. The idea that light has a pressure was put forth by Kepler to explain the deflection of comet tails from the Sun. However, this idea raised a lot of controversy since the existence of light pressure could not be proved for quite a long time. Supporters of the wave theory maintained for a long time that light pressure does not exist, and used the absence of experimental evidence of light pressure as an argument against the corpuscular theory which uniequivocally supports the existence of light pressure. However, the electromagnetic theory of light does predict the existence of light pressure. If an electromagnetic wave is incident normally on the plane surface of a solid and is completely absorbed by it, then in accordance with (3.11), a momentum G is imparted to 1 m^2 of the solid surface in 1 s. Hence the light pressure is

$$\boxed{p = cG = S/c.} \tag{3.12}$$

In the case of total reflection a momentum imparted to the solid surface is twice as large, and hence a pressure is twice as high. The pressure in the case of a partial absorption of energy flux can be calculated. If the flux density of the absorbed energy is $S_{\text{abs}} = \alpha S$, then in accordance with the law of energy conservation the flux density of the scattered energy will be $S_{\text{scat}} = (1 - \alpha) S$, and hence the pressure can be expressed in the form $p = \alpha S/c + 2(1 - \alpha) S/c = (2 - \alpha) S/c$. When the energy flux is incident at an angle other than 90°, we must take into consideration just the normal component of the energy flux density. Under ordinary conditions, the light pressure is very low. For example, the flux density of solar radiation on the Earth orbit is about $S = 1400$ W/m^2. Hence $p = 1400/(3 \times 10^8)$ Pa $= 0.5 \times 10^{-5}$ Pa.

Considering that the atmospheric pressure is about 10^5 Pa, we conclude that the light pressure on the Earth is lower than the atmospheric pressure by a factor of about 10^{10}. Hence it was very difficult to detect this pressure. The first step in the right direction was taken by Crookes (1832-1919) who used a torsion balance. The foils of the balance were exposed to light. The force acting on the foil could be estimated from the twisting of the thread. However, Crookes was unable to measure the light pressure in view of the radiometric effect. Lebedev (1866-1912) was the first to measure the light pressure in 1900. He countered the radiometric forces by creating a high vacuum in the vessel containing the torsion balance. As a result, the force of the light pressure became the dominating factor in twisting the thread of the torsion balance and was measured.

The reliability of the elimination of radiometric force in this manner was experimentally verified as follows. If the surface at which the ray is directed is reflecting, the light pressure on this surface will be twice that for a totally absorbing surface, while the radiometric effect will be less pronounced. If, on the other hand, the surface totally absorbs radiation, the light pressure on it will be half that for a totally reflecting surface, and the radiometric effect will be more pronounced. Lebedev indeed observed a twofold increase in the light pressure when the reflecting foil of the torsion balance was exposed to light, compared to the case when the absorbing (black) foil was exposed to light. This proves that the radiometric effect was indeed eliminated in Lebedev's experiments.

?

What do we mean by the existence of the instantaneous values of quantities if they cannot be measured?
Enumerate the basic parameters of a Gaussian beam.
What is the origin of radiometric forces and how were they eliminated in Lebedev's experiments?
What are the main factors responsible for the compression of a target in laser fusion? Can it be stated that all these factors lead to the light pressure?
Write down the formulas for the transformation of the energy and momentum of a plane wave train upon a transition from one reference system to another.

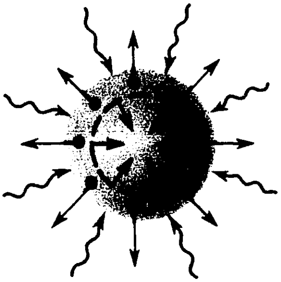


Fig. 9. Realization of laser fusion.

EFFECT OF LIGHT PRESSURE ON SMALL PARTICLES. The light pressure is proportional to the area, i.e. squared linear dimensions of a particle, while the particle mass is proportional to the volume, i.e. the third power of linear dimensions. This means that for a fixed density, the force per unit mass due to the light pressure is inversely proportional to the linear dimensions of the particle, i.e. increases with decreasing particle size. Suppose, for example, that the density of a spherical particle is $\rho = 10^3$ kg/m^3 and the radius $r = 10^{-5}$ m. The mass of the particle will then be $m = (4/3)\pi r^3 \rho = 4 \times 10^{-12}$ kg. For a solar radiation flux $S = 1400$ W/m^2 on the Earth orbit, the light pressure $p = S/c = 0.5$ µPa. Hence the force acting on a particle that completely absorbs radiation is given by $F = p\pi r^2 = 1.5 \times 10^{-15}$ N. This force imparts an acceleration $a = F/m = 0.4 \times 10^{-3}$ m/s^2 to the particle. The acceleration imparted by the Sun to a particle in the Earth orbit is $a_S = v^2/R = (30 \times 10^3)^2/(15 \times 10^{10})$ m/s$^2 = 0.6 \times 10^{-2}$ m/s^2, which is about $(a_S/a) \simeq 15$ times larger than the acceleration caused by the light pressure. For a particle whose linear dimensions are smaller by a factor of 15, the force due to the light pressure will practically neutralize the attractive force of the Sun. Since the radiation flux density as well as the gravitational force decrease with distance according to the same law (in inverse proportion to the square of the distance from the Sun), the force of attraction of the Sun and the force due to the light pressure are equal in magnitude for all distances and hence a particle moves in the gravitational field of the Sun as if this field were absent. The most clear manifestation of the light pressure is the orientation of the comet tails as they pass near the Sun ("near" on the astronomical scale of distances).

LASER FUSION. This term is employed for describing a controlled fusion reaction initiated by laser radiation. For this purpose, light nuclei (deuterium or tritium) have to be brought together to such small distances where a nuclear fusion reaction can take place. However, this requires the overcoming of the Coulomb repulsive force between the positively charged nuclei, and this means that a considerable amount of kinetic energy has to be supplied to these nuclei (the material has to be heated).

The laser fusion is based on the following idea. Suppose that a small ball containing a nuclear fuel (deuteron or tritium nuclei) is subjected simultaneously to powerful short laser pulses from all sides (Fig. 9, undulating arrows). The energy and momentum of this radiation are imparted to the matter in the surface layer of the ball which gets heated and acquires a directed motion towards the centre. The energy and momentum of the matter are then transferred to the inner part of the ball which becomes strongly compressed and gets heated. The particles in the outer layers acquire a velocity from the centre (and are as if "vapourized" from the surface). Thus, the matter in the inner part of the ball is compressed very strongly, and this is accompanied by an enormous rise in temperature. The matter from the outer layers of the ball flies off in all directions with very high velocities (Fig. 9, straight arrows). If the density and temperature of the compressed matter in the ball acquire values sufficient for inducing nuclear fusion, a small thermonuclear explosion like that of a hydrogen bomb takes place. The energy released in this process is mainly converted into the kinetic energy of the fusion reaction products and, in principle, can be transformed into other forms of energy and used gainfully.

To achieve a fusion reaction, the material of the ball must be compressed to a few thousands of its original size, and the temperature must be increased to tens of millions degrees. This is possible only with the help of highly powerful laser radiation pulses.

However, the attainment of nuclear fusion is a simple task only from a theoretical point of view. Its technical implementation is extremely complicated and involves a great deal of research. It turned out that the compression and heating of matter in spherical bodies is a very complex process and depends on many factors. The problem is further complicated by the fact that the attainment of a miniature thermonuclear explosion in spherical objects does not mean that controlled nuclear fusion reactions can be realized in practice. The net energy balance for the operation of the installation must be positive, i.e. the energy released by the system in a form that can be used purposefully must be higher than the energy used for running the installation.

TRANSFORMATION OF AMPLITUDE AND NORMAL VECTOR OF A PLANE ELECTROMAGNETIC WAVE. Suppose that the wave vector of a plane wave lies in the XY-plane. We can write the following expressions for the electric field strength and magnetic induction:

$$E_x = -E_0 n_y e^{i\varphi}, \quad E_y = E_0 n_x e^{i\varphi}, \quad B_z = (E_0/c)\, e^{i\varphi}, \tag{3.13}$$

where φ and E_0 are the phase and amplitude of the plane wave. Since the phase of the wave is invariant to the Lorentz transformation, we obtain from the field vector transformation formulas

$$E_0 = E_0'(1 + \beta n_x)/\sqrt{1-\beta^2}, \tag{3.14a}$$

$$n_x E_0 = E_0'(n_x' + \beta)/\sqrt{1-\beta^2}, \tag{3.14b}$$

$$n_y E_0 = n'_y E'_0. \tag{3.14c}$$

As before, the primed quantities correspond to the coordinate system K' moving with a velocity v in the positive direction of the X-axis in the system K. The X- and X'-axes in the two coordinate systems are made to coincide.

Dividing Eqs. (3.14b) and (3.14c) by (3.14a) termwise, we get the following transformation formulas for the projections of the unit normal vector of the plane wave:

$$n_x = (n'_x + \beta)/(1 + \beta n'_x), \tag{3.15a}$$

$$n_y = n'_y \sqrt{1 - \beta^2}/(1 + \beta n'_x). \tag{3.15b}$$

Dividing Eq. (3.14b) by the first equation in (2.61) termwise, we obtain

$$\boxed{E_0/\omega = E'_0/\omega'.} \tag{3.16}$$

This relation between the amplitude and frequency of a plane wave is invariant to the Lorentz transformation.

ENERGY OF A PLANE WAVE TRAIN. According to (3.4), the energy contained in volume V is given by

$$W = (1/2)\varepsilon_0 E_0^2 V. \tag{3.17}$$

The volume occupied by a wave train is moving at a velocity c. Hence we can neither attach a system of coordinates to this volume, nor speak of the magnitude of this volume in the state of rest. However, we can find the transformation formula for the volume of a wave train. We introduce a certain auxiliary volume V_0 which moves in the coordinate system K' with a velocity u'_x, and in the system K with a velocity u_x. Taking into account the relativistic contraction of lengths, we can write

$$V' = V_0 \sqrt{1 - u_x'^2/c^2}, \quad V = V_0 \sqrt{1 - u_x^2/c^2} \tag{3.18}$$

and hence

$$\frac{V'}{V} = \frac{\sqrt{1 - u_x'^2/c^2}}{\sqrt{1 - u_x^2/c^2}}. \tag{3.19}$$

Combining the last formula with the velocity summation relation in the theory of relativity, we obtain the following relation:

$$V' = \frac{V\sqrt{1 - \beta^2}}{1 - v u_x/c^2}. \tag{3.20}$$

Replacing u_x in this relation with the help of the formula

$$u_x = u n_x, \tag{3.21}$$

where n_x is the X-projection of the unit vector characterizing the direction of velocity u, we obtain

$$V' = \frac{V\sqrt{1-\beta^2}}{1 - vun_x/c^2}. \tag{3.22}$$

As $u \to c$, we obtain the transformation relation for the volumes of a wave train moving with the velocity of light:

$$V = \frac{V\sqrt{1-\beta^2}}{1 - \beta n_x}. \tag{3.23}$$

Using the relativity principle, we can write formula (3.14a) as follows:

$$E_0' = E_0(1 - \beta n_x)/\sqrt{1-\beta^2}. \tag{3.24}$$

Comparing (3.23), (3.24), (3.17) and (2.62), we obtain

$$\boxed{W'/\omega' = W/\omega, \qquad V'\omega' = V\omega.} \tag{3.25}$$

This means that the energy of a plane wave train is directly proportional to the frequency. This result is used in the quantum theory of light.

MOMENTUM OF A PLANE WAVE TRAIN. This quantity is equal to

$$\mathbf{G} = (1/c^2)\int \mathbf{E} \times \mathbf{H}\,dV = \mathbf{n}(\varepsilon_0/c)\int E^2 dV = \mathbf{n}W/c, \tag{3.26}$$

where n is a unit vector in the direction of wave propagation. All the quantities in (3.26) have instantaneous values (i.e. averaging over the oscillation period has not been envisaged). The following relations have been used for deriving formula (3.26):

$$EH/c^2 = \varepsilon_0 E^2/c = w/c. \tag{3.27}$$

With the help of (3.25), we can write (3.26) in the following form:

$$\mathbf{G} = \mathbf{n}\omega\alpha, \quad W = \omega c\alpha, \tag{3.28}$$

where $\alpha = W/(\omega c)$ is a constant quantity in accordance with (3.25). From (2.60b), we find that the set of quantities

$$(\omega n_x,\ \omega n_y,\ \omega n_z,\ -i\omega) \tag{3.29}$$

forms a four-dimensional vector. Hence it follows from (3.28) that the set of quantities

$$(G_x,\ G_y,\ G_z,\ -iW/c), \tag{3.30}$$

characterizing the total momentum and total energy of a plane wave train also forms a four-dimensional vector.

!

The instant values of quantities cannot be measured experimentally at points in space. Only the mean values of quantities over a certain interval of time and a certain region of space are always measured in experiments.

At present, it is not possible to measure experimentally the mean values of quantities over time intervals shorter than the period of light vibrations. Only the values averaged over many periods are always measured. The experimentally measured mean value of the electric field strength of a light wave is always equal to zero. The simplest quantities whose mean value is not zero are proportional to the square of the field strength. The most important among these is the volume energy density which is connected with the energy flux density and the volume momentum density through a simple relation.

The light pressure is a manifestation of the volume density of the wave momentum and the law of momentum conservation when a wave interacts with matter.

Example 3.1. A ruby laser emits linearly polarized light pulses with a Gaussian distribution of amplitudes over the circular beam cross section. The pulse energy and duration, and the effective beam radius are respectively equal to 1 J, 100 μs and 5 mm. Assuming that the instantaneous emissive power is constant over the duration of a pulse, determine its value, the energy flux density on the beam axis, the amplitude of the electric field strength and the amplitudes of the magnetic induction and field strength at a distance of 2 mm from the beam axis.

Since the instantaneous value of the emissive power is constant, the mean power in a pulse is equal to the instantaneous value: $P = W/\tau = 10^4$ W [see (3.10)]. The energy flux density on the beam axis is [see (3.8)] $S_0 = P/(\pi r_0^2) = 1.27 \times 10^8$ W/m^2. The amplitude of the electric field strength on the axis will then be equal to [see (3.4)] $E_0(0) = [2S_0/(c\varepsilon_0)]^{1/2} = 3.1 \times 10^5$ V/m. The electric field strength at a distance of 2 mm from the axis is [see (3.6)]

$$E_0(2 \times 10^{-3}\ \text{m}) = E_0(0) \exp[-r^2/(2r_0^2)] = 2.86 \times 10^5\ \text{V/m}.$$

Hence we obtain the following values for the amplitudes of the magnetic induction and magnetic field strength: 0.95×10^{-3} T and 7.56×10^2 A/m.

SEC. 4. SUPERPOSITION OF ELECTROMAGNETIC WAVES

We trace the transition from the superposition principle for the electromagnetic field strength of a wave to the superposition principle for an electromagnetic wave.

SUPERPOSITION OF FIELD VECTORS OF A WAVE. The electric field strength and the magnetic induction are respectively equal to the sum of the field strengths and magnetic inductions at a given point irrespective of their origin. In particular, these fields may belong to plane electromagnetic waves of all possible frequencies and directions of propagation. However, the set of electric and magnetic fields obtained as a result of summation of individual fields generally does not form a travelling electromagnetic wave even if the components of a field belong to travelling electromagnetic waves.

SUPERPOSITION OF TRAVELLING MONOCHROMATIC ELECTROMAGNETIC PLANE WAVES. Suppose that we have two waves with identical wave vectors $\mathbf{k}$ and the same frequency ω. Let the fields of these waves be described by $\mathbf{E}_1$, $\mathbf{B}_1$ and $\mathbf{E}_2$, $\mathbf{B}_2$. For these waves, formulas (2.53)-(2.56) assume the following form:

$$\begin{aligned} -\mathbf{k}\times\mathbf{B}_1 &= \omega\mu_0\varepsilon_0\mathbf{E}_1, & -\mathbf{k}\times\mathbf{B}_2 &= \omega\mu_0\varepsilon_0\mathbf{E}_2, \\ \mathbf{k}\times\mathbf{E}_1 &= \omega\mathbf{B}_1, & \mathbf{k}\times\mathbf{E}_2 &= \omega\mathbf{B}_2, \\ \mathbf{k}\cdot\mathbf{B}_1 &= 0, & \mathbf{k}\cdot\mathbf{B}_2 &= 0, \\ \mathbf{k}\cdot\mathbf{E}_1 &= 0, & \mathbf{k}\cdot\mathbf{E}_2 &= 0, \end{aligned} \tag{4.1}$$

while formula (2.57) leads to the equalities

$$E_1 = cB_1, \quad E_2 = cB_2. \tag{4.2}$$

Adding the left- and right-hand sides of Eqs. (4.1) and (4.2) termwise and using the notation

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2, \quad \mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2, \tag{4.3}$$

we arrive at the relations

$$-\mathbf{k}\times\mathbf{B} = \omega\mu_0\varepsilon_0\mathbf{E},\ \mathbf{k}\times\mathbf{E} = \omega\mathbf{B},\ \mathbf{k}\cdot\mathbf{B} = 0,\ \mathbf{k}\cdot\mathbf{E} = 0, \tag{4.4}$$
$$E = cB,$$

which prove that the field described by the vectors $\mathbf{E}$ and $\mathbf{B}$ is a plane monochromatic travelling wave with a wave vector $\mathbf{k}$ and frequency ω. Thus, we have proved that **two plane monochromatic travelling waves having the same frequency and propagating in the same direction can be added to form a plane monochromatic wave having the same frequency and propagating in the same direction. If the component waves have different frequencies or different directions, their addition will not give a plane monochromatic travelling wave.**

BEATS. Let us consider the summation of two monochromatic waves having frequencies ω_1 and ω_2 and propagating in the same direction. The vectors $\mathbf{E}$ of these waves are collinear. For the sake of definiteness, we make the Z-axis coincide with the direction of the wave propagation and direct the X-axis along the vector $\mathbf{E}$ of the wave, i. e. we assume that $\mathbf{E} = (E_x, 0, 0)$ and $\mathbf{B} = (0, B_y, 0)$. In order to simplify the analysis, we shall follow the variation of vector $\mathbf{E}$ since the behaviour of vector $\mathbf{B}$ depends on that of vector $\mathbf{E}$ and can be determined from the relations between the plane wave vectors. For the sake of simplicity, we also assume that the electric field strengths of the component waves have the same amplitude:

$$E_{1x} = E_0\cos(\omega_1 t - k_1 z),\ E_{2x} = E_0\cos(\omega_2 t - k_2 z). \tag{4.5}$$

According to the superposition principle, we have

$$\begin{aligned} E = E_x = E_{1x} + E_{2x} &= E_0\cos(\omega_1 t - k_1 z) + E_0\cos(\omega_2 t - k_2 z) \\ &= 2E_0\cos\left(\frac{\omega_1-\omega_2}{2}t - \frac{k_1-k_2}{2}z\right) \\ &\quad\times\cos\left(\frac{\omega_1+\omega_2}{2}t - \frac{k_1+k_2}{2}z\right), \end{aligned} \tag{4.6}$$

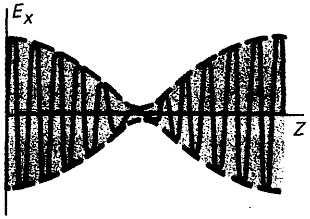


Fig. 10. Beats.

where we have used the cosine summation formula. Considering that $k_1 = \omega_1/c$, $k_2 = \omega_2/c$, we can express (4.6) as follows:

$$E = 2E_0 \cos [(\omega_1 - \omega_2)(t - z/c)/2] \times \cos [(\omega_1 + \omega_2)(t - z/c)/2]. \tag{4.7}$$

This relation shows that the resultant electromagnetic field propagates without attenuation in the positive direction of the Z-axis with the velocity c (this is confirmed by the presence of the combination $t - z/c$ in the argument of the function). In this respect, we are speaking of a travelling, though not monochromatic, wave. Considering that the relation $|\omega_1 - \omega_2| \ll \omega_1 + \omega_2$ is always satisfied in the optical range, we can provide the following visual interpretation for such a wave: the factor $2E_0 \cos [(\omega_1 - \omega_2)(t - z/c)/2]$ represents the varying amplitude of the harmonic wave with a frequency $(\omega_1 + \omega_2)/2$ [the last factor in (4.7)]. Thus, at a certain fixed instant of time, the wave has the form shown in Fig. 10. The solid line shows the oscillations with a frequency $(\omega_1 + \omega_2)/2$, while the dashed line shows the envelope of the amplitudes of oscillations varying from zero to their peak value $2E_0$. If the amplitudes E_{01} and E_{02} of the fields of component waves are not equal, the amplitude of the resultant wave varies from $E_{01} + E_{02}$ to $|E_{01} - E_{02}|$.

Harmonic oscillations with a slowly varying amplitude are called beats. The concept of "slowly varying amplitude" is referred to the fundamental harmonic oscillation: the amplitude does not vary significantly over several periods of the fundamental harmonic oscillation (Fig. 10).

Since by definition the oscillation amplitude is the positive value of the maximum deviation from the equilibrium position, we conclude from (4.7) that the beat frequency is equal to

$$\boxed{\Omega = |\omega_1 - \omega_2|.} \tag{4.8}$$

?

Which property of electromagnetic waves ensures the observance of the superposition principle for them as a direct corollary of the principle of superposition for electric field strength and the magnetic induction?
In which regions of space is the energy transported in a standing wave? Consider the problem of energy transfer in the presence of beats.

STANDING WAVES. Let us consider the superposition of two monochromatic waves of the same frequency moving towards each other. We shall assume the electric field vectors of these waves to be collinear and oscillating with the same amplitude. As before, we direct the Z-axis along the direction of the wave propagation and make the X-axis coincide with the direction of the vectors E of the two waves. We have

$$E_1 = E_{1x}(z, t) = E_0 \cos(\omega t - kz), \tag{4.9}$$

$$E_2 = E_{2x}(z, t) = E_0 \cos(\omega t + kz + \delta), \tag{4.10}$$

where the positive sign of kz in (4.10) means that the wave with E_2 propagates in the negative direction of the Z-axis, and δ is the phase shift.

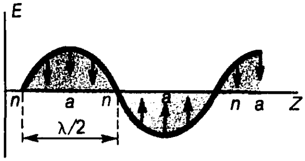


Fig. 11. A standing wave of strength *E*.

The superposition of these two travelling waves leads to the formation of a wave whose field strength is given by

$$E = E_1 + E_2 = 2E_0 \cos(kz + \delta/2) \cos(\omega t + \delta/2). \qquad (4.11)$$

It is obvious that this is not a travelling wave since it does not contain the characteristic factor $t \pm z/c$. Except for the sign, the factor $2E_0 \cos(kz + \delta/2)$ can be considered as the amplitude of field strength oscillations at the point z. It varies according to a harmonic law from point to point. The field strength at all points varies with the same frequency and in the same phase [the factor $\cos(\omega t + \delta/2)$]. Such a wave is called a standing wave (Fig. 11). At points on the Z-axis where the condition $\cos(kz + \delta/2) = 0$ is satisfied, the field strength E is identically equal to zero. These points are called nodes (letter n in the figure). The field strength oscillations have the maximum amplitude at points where $\cos(kz + \delta/2) = \pm 1$. These points are called antinodes (letter a in the figure). Obviously, the distance Δz between nodes (or antinodes) is determined from the condition $k\Delta z = \pi$ and is equal to $\Delta z = \lambda/2$, i.e. half the length of the travelling wave. For nonzero values of the electric field strength, a standing wave has the form shown in Fig. 11. It can be seen that the instantaneous pattern of a standing wave coincides with that of a travelling wave. However, there is a fundamental difference between travelling and standing waves. The oscillations of field strength at all points of a standing wave are in the same phase at a particular instant of time, while the field oscillations at different points of a travelling wave are not in phase. In particular, at a certain instant of time, the field strength E at all points on the Z-axis is equal to zero for a standing wave [when $\cos(\omega t + \delta/2) = 0$].

The magnetic induction **B** of the wave fields is also added in accordance with the superposition principle for fields. Since the vectors **E**, **B** and **k** of a plane electromagnetic wave form a right-handed triplet of vectors (see Fig. 6), the vectors $\mathbf{B}_1$ and $\mathbf{B}_2$ of the waves propagating in the positive and negative directions of the Z-axis can be written in the following form in accordance with (4.9) and (4.10):

$$B_1 = B_{1y}(z, t) = E_{1x}/c = (E_0/c) \cos(\omega t - kz), \qquad (4.12)$$

$$B_2 = B_{2y}(z, t) = -E_{2x}/c = -(E_0/c) \cos(\omega t + kz + \delta) \qquad (4.13)$$

(the minus sign on the right-hand side indicated that the right-handed triplet of vectors is formed by vector **E** in the positive direction of the X-axis, vector **B** in the negative direction of the Y-axis and vector **k** in the negative direction of

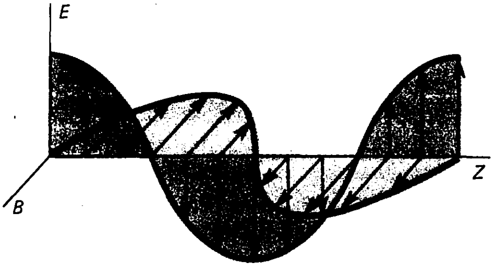


Fig. 12. A standing electromagnetic wave.

the Z-axis). The magnetic induction of the resultant wave is

$$B = B_y = B_1 + B_2 = (2E_0/c)\sin(kz + \delta/2)\sin(\omega t + \delta/2), \quad (4.14)$$

i. e. has a form similar to formula (4.11) but with sine in the place of cosine. This means that the vector **B** also forms a standing wave whose nodes coincide with the antinodes of the standing wave for **E**. The vectors **E** and **B** lie in mutually orthogonal planes. A comparison of (4.11) and (4.14) also shows that the electric and magnetic field oscillations of a standing electromagnetic wave differ in phase by a quarter of the period of oscillations. This means that if, for example, the electric field strength of a standing wave assumes its peak value, the magnetic induction at this instant is equal to zero; if the electric field strength of the standing wave reaches half its peak value at a certain instant, the magnetic induction at this instant also has half its peak value, the only difference being that the electric field strength is, say, in the ascending phase of its magnitude, while the magnetic induction is in the descending phase.

Figure 12 shows a standing wave at an instant when neither the electric field strength nor the magnetic induction has reached its peak value.

ENERGY TRANSFORMATION IN A STANDING ELECTROMAGNETIC WAVE. The energy flux density of waves is described by Poynting's vector (3.1). Hence there is no energy flux at points where **E** or **B** is equal to zero. This means that there is no energy flux through nodes and antinodes in a standing electromagnetic wave, since the antinode of the electric field strength coincides with the node of the magnetic induction, and vice versa. Hence with the passage of time, energy moves between neighbouring nodes and antinodes, being transformed from the magnetic field energy into the electric field energy and vice versa. Using formulas (4.11) and

(4.14) and the expression

$$\underline{\overline{w = (1/2)(\mathbf{E}\cdot\mathbf{D} + \mathbf{H}\cdot\mathbf{B})}} \tag{4.15}$$

for the volume density of the electromagnetic field energy, we can verify that the energy of a standing wave between successive nodes and antinodes remains constant with the passage of time, thus proving the statement on the mutual conversion of electric and magnetic field energies for a standing wave.

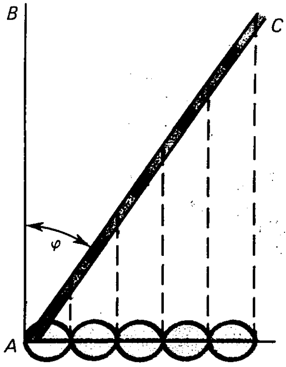


Fig. 13. Wiener experiment to prove the electromagnetic nature of light.

!

The principle of superposition for electric and magnetic fields does not automatically lead to its validity for waves. A special proof is required for the superposition principle for electromagnetic waves.
An electromagnetic wave and the volume density of its energy cannot exist without motion. Motion is the way of their existence. Hence, even in a standing electromagnetic wave, the electric field strength, as well as the energy density per unit volume change continuously.
Wiener's experiments are important not only because they provide a direct experimental verification of the electromagnetic nature of light, but also because they confirm the fact that the electric field strength of the wave acts on a photographic emulsion.

EXPERIMENTAL PROOF OF THE ELECTROMAGNETIC NATURE OF LIGHT. Standing light waves are obtained as a result of superposition of an incident light beam and the beam reflected from a certain surface. The laws of reflection will be considered somewhat later (see Secs. 16,20). For the present, we simply observe that to change the direction of motion of a wave upon reflection, the mutual orientation of vectors E and **B** should be changed in such a way that the vector product E × **B** is directed along the reflected ray, while its direction coincides with the incident ray before reflection. Hence, either the vector E of the incident wave or the vector **B** must reverse direction as a result of reflection.

If a standing light wave is created in a photographic emulsion, the antinodes of the standing wave should become the darkest spots. If the photographic action of the electric and magnetic vectors is identical, their antinodes will be darkened to the same extent. If this is not so, the darkening of the two types of antinodes will be different. Thus, an analysis of standing waves in photographic emulsion helps in a verification of the electromagnetic theory of light and in establishing the extent of the photographic activity of the wave vectors. Such experiments were carried out for the first time in 1890 by Wiener who obtained a system of standing waves in air by reflection from a metallic mirror. Since the distance between antinodes is very small (of the order of 0.3 μm), Wiener investigated the darkening in a thin light-sensitive layer AC (about $\lambda/20$ thick) at a small angle φ to the mirror surface AB (Fig. 13). If the distance between antinodes along the normal to the mirror surface is $\lambda/2$, this distance will be equal to $d = \lambda/(2\sin\varphi)$ in a thin sensitive film inclined to the normal. This quantity can be made quite large and measured by choosing small values of the angle φ.

Wiener's experiments showed that the photographic action is due to the electric field E and not to the magnetic induction B. The predictions of the electromagnetic theory of light were fully confirmed in Wiener's experiments.

SEC. 5. POLARIZATION OF ELECTROMAGNETIC WAVES

It is shown that electromagnetic waves have two independent polarizations. Various representations of waves are analyzed.

POLARIZATION. For longitudinal waves, all directions perpendicular to the direction of wave propagation are equivalent. This is not true in the case of transverse waves. Electromagnetic waves are transverse, and hence their properties depend on the orientation of vectors **E** and **B**, which is associated with the concept of polarization.

LINEAR POLARIZATION. If in the process of wave propagation the vector **E** remains in the same plane parallel to the direction of propagation, the waves are called linearly polarized (see Fig. 7). The plane in which the vectors **B** and **k** (vector **B** and not **E**!) lie is called the plane of polarization. However, this concept is rarely used at present. More often, we speak of the plane of oscillations of vector **E**.

SUPERPOSITION OF LINEARLY POLARIZED WAVES. Let us consider the superposition of two linearly polarized waves having the same frequency and propagating in the same direction. For the sake of definiteness, we shall assume that the oscillations of **E** of the first wave lie in the XZ-plane and of the second wave, in the YZ-plane (Fig. 14). Then we can write

$$E_{1x}(z,t) = E_{10}\sin(\omega t - kz), \quad E_{1y} = E_{1z} = 0, \tag{5.1}$$

$$E_{2y}(z,t) = E_{20}\sin(\omega t - kz + \delta), \quad E_{2x} = E_{2z} = 0, \tag{5.2}$$

where δ is the phase shift between the oscillations.

Let us analyze the strength $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$ of the electric field of the resultant wave in a plane perpendicular to the direction of propagation of waves for a fixed value of z. With the passage of time, the tip of the vector **E** describes a certain closed curve in the XY-plane. Let us find the equation of this curve. We can rewrite (5.2) in the form

$$E_y = E_{20}\sin(\omega t - kz)\cos\delta + E_{20}\cos(\omega t - kz)\sin\delta$$

and eliminate $\sin(\omega t - kz)$ and $\cos(\omega t - kz)$ from this equation with the help of (5.1):

$$E_y = E_{20}(E_x/E_{10})\cos\delta + E_{20}\sin\delta\sqrt{1 - E_x^2/E_{10}^2}. \tag{5.3}$$

It should be recalled that the amplitudes E_{10} and E_{20} are assumed to be positive numbers. Transposing the first term on the right-hand side to the left, squaring both sides and dividing them by E_{20}^2, we can reduce the equation to the following form after regrouping the terms:

$$\frac{E_x^2}{E_{10}^2} + \frac{E_y^2}{E_{20}^2} - 2\frac{E_x}{E_{10}}\frac{E_y}{E_{20}}\cos\delta = \sin^2\delta. \tag{5.4}$$

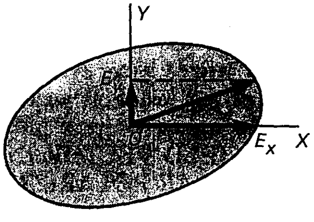


Fig. 14. Superposition of linearly polarized plane waves, resulting in an elliptically polarized wave.

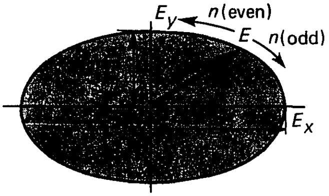


Fig. 15. Elliptical polarization for the case when the principal axes of the ellipse are collinear with the vibrations of E.

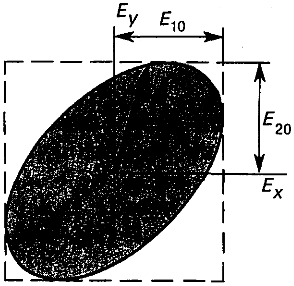


Fig. 16. Elliptical polarization in the general case.

?

For elliptic polarization, does the orientation of the principal axes of the ellipse depend on the phase difference between the waves that are linearly polarized in mutually perpendicular directions and whose superposition results in the formation of an elliptically polarized wave?
Is it possible to obtain a circularly polarized wave by a superposition of linear waves of different amplitudes? Is it possible to obtain an elliptically polarized wave by a superposition of linearly polarized waves having the same amplitude? On what factors does the direction of circular polarization depend?

We shall consider different cases described by this equation.

ELLIPTICAL AND CIRCULAR POLARIZATIONS. If $\cos\delta = 0$ and $\sin\delta = \pm 1$, Eq. (5.4) assumes the form

$$E_x^2/E_{10}^2 + E_y^2/E_{20}^2 = 1. \tag{5.5}$$

For $E_{10} \neq E_{20}$, this is the equation of an ellipse with its centre at the origin of coordinates and its axes along the coordinate axes. The semiaxes of the ellipse (Fig. 15) are E_{10} (along the X-axis) and E_{20} (along the Y-axis). The condition $\cos\delta = 0$ is satisfied for

$$\delta = \pi/2 + n\pi \quad (n = 0, \pm 1, \pm 2, \ldots). \tag{5.6}$$

Then Eqs. (5.1) and (5.2) assume the following form ($z = 0$):

$$E_x = E_{10} \sin\omega t, \tag{5.7}$$

$$E_y = E_{20} \sin(\omega t + \pi/2 + n\pi) = (-1)^{n+1} E_{20} \cos\omega t. \tag{5.8}$$

It can be seen that the tip of vector E rotates in the clockwise direction for odd values of n and in the counterclockwise direction for even n. In the former case, we speak of a right-hand elliptically polarized wave, while in the latter case a left-hand elliptically polarized wave is meant. The rotation of vector E is observed from the side towards which the wave propagates (in Fig. 15 the Z-axis is directed towards us).

For $E_{10} = E_{20}$, the ellipse becomes a circle. The corresponding wave is said to be polarized along a circle, or just circularly polarized. As in the case of elliptically polarized waves, there exist waves with left- or right-hand circular polarization.

For $\cos\delta \neq 0$, Eq. (5.4) also describes an ellipse, but its principal axes no longer coincide with the coordinate axes. It can be seen from Eqs. (5.1) and (5.2) that the maximum and minimum values of the components E_y and E_x are equal to $\pm E_{20}$ and $\pm E_{10}$. Hence the ellipse is inscribed in a rectangle with sides $2E_{10}$ and $2E_{20}$, and with its centre at the origin of coordinates (Fig. 16). The orientation of the ellipse and its parameters depend on δ. In particular, it should be noted that for $\cos\delta \neq 0$ we obtain an elliptically polarized wave even for $E_{10} = E_{20}$. The direction of rotation of the resultant vector E is determined by the value of δ.

VARIATION OF THE ELECTRIC FIELD VECTOR IN SPACE FOR ELLIPTICAL AND CIRCULAR POLARIZATION. At a fixed moment of time, the set of points formed by the tip of the vector E lies on a helical line. Figure 17 shows the spatial variation of the vector E for a circular polarization of the wave. For a fixed t, the elliptically polarized wave can be imagined as follows: a helical line is drawn on the surface of

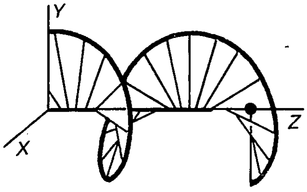


Fig. 17. Variation of vector E in space for circularly polarized waves at a fixed instant of time.

a straight elliptical cylinder. The beginning of the vector E lies on the cylinder axis, while its tip lies on the helical line, the vector itself being perpendicular to the axis everywhere.

DEGENERATE CASE OF ELLIPTICAL POLARIZATION. For $\cos\delta = \pm 1$ and $\sin\delta = 0$, expression (5.4) is transformed into the equality

$$\frac{E_x^2}{E_{10}^2} + \frac{E_y^2}{E_{20}^2} \mp 2\frac{E_x}{E_{10}}\frac{E_y}{E_{20}} = \left(\frac{E_x}{E_{10}} \mp \frac{E_y}{E_{20}}\right)^2 = 0.$$

For $\cos\delta = +1$ and $\cos\delta = -1$, this relation describes the straight lines

$$E_x/E_{10} - E_y/E_{20} = 0 \qquad (5.9a)$$

and

$$E_x/E_{10} + E_y/E_{20} = 0. \qquad (5.9b)$$

The tip of the resultant vector E moves along the corresponding straight line (Fig. 18*a* and *b*). We obtain a linearly polarized wave which is the limiting case of an elliptically polarized wave (see Fig. 16) when one of the semiaxes of the ellipse is equal to zero. For $\cos\delta = +1$, the linear oscillation of the resultant vector E occurs in the first and third quadrants, while for $\cos\delta = -1$, the oscillation takes place in the second and fourth quadrants.

NUMBER OF INDEPENDENT POLARIZATIONS. The above description shows that an electromagnetic wave with any polarization can be represented as a superposition of two linearly polarized waves whose electric vectors oscillate in mutually orthogonal planes. Hence it can be stated that electromagnetic waves have two independent states of polarization.

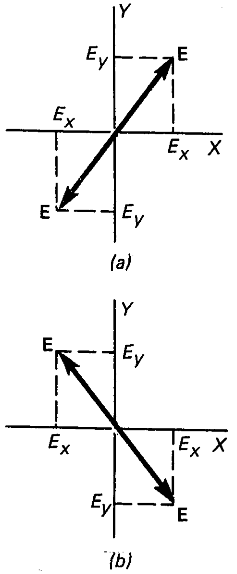


Fig. 18. Degenerate case of summation of two linearly polarized waves.

LINEARLY POLARIZED WAVE AS A SUPERPOSITION OF CIRCULARLY POLARIZED WAVES. Let us consider the superposition of a left- and a right-hand polarized wave. Suppose that for a fixed *z*, the electric field vectors E_1 and E_2 of the waves are described by the relations (Fig. 19).

$$E_{1x} = E_0 \cos\omega t, \quad E_{1y} = E_0 \sin\omega t, \qquad (5.10)$$

$$E_{2x} = E_0 \cos\omega t, \quad E_{2y} = -E_0 \sin\omega t. \qquad (5.11)$$

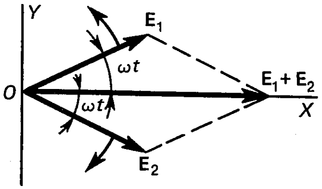


Fig. 19. Formation of a linearly polarized wave by superposition of two circularly polarized waves.

The first wave has left-hand polarization while the second one has right-hand polarization. Both waves have identical amplitudes. As a result of superposition of the waves, we obtain a wave with the following projections of the electric field strength:

$$E_x = E_{1x} + E_{2x} = 2E_0 \cos\omega t, \qquad (5.12)$$

$$E_y = E_{1y} + E_{2y} = 0. \qquad (5.13)$$

In other words, we obtain a linearly polarized wave. In this case, the line along which the vector E oscillates coincides with the X-axis. If there is a phase shift between the oscillations [see (5.10) and (5.11)], the line of oscillations of the resultant electric field forms an angle with the X-axis, whose magnitude is determined by the phase shift between the component waves.

Example 5.1. Find the characteristics of the wave obtained by superposition of two waves having the same amplitude and right- and left-hand circular polarizations, if the initial phase difference between the waves is δ.

The electric field strengths of the left- and right-hand polarized waves can be written in the form

$$E_{1x} = E_0 \cos \omega t, \quad E_{1y} = E_0 \sin \omega t,$$
$$E_{2x} = E_0 \cos(-\omega t + \delta), \quad E_{2y} = E_0 \sin(-\omega t + \delta). \tag{5.14}$$

It is convenient to add the waves described by formulas (5.14) if we represent them in the form of complex vectors:

$$E_1 = E_{1x} + iE_{1y} = E_0 e^{i\omega t}, \quad E_2 = E_{2x} + iE_{2y} = E_0 e^{-i(\omega t - \delta)}. \tag{5.15}$$

!

It follows from the principle of superposition that any linearly polarized wave can be presented as the sum of two waves with mutually perpendicular planes of linear polarization. If a phase difference arises between these waves, the resultant wave is polarized either elliptically or circularly, depending on the phase difference.

The superposition of these waves gives

$$E = E_1 + E_2 = E_0 e^{i\delta/2}[e^{i(\omega t - \delta/2)} + e^{-i(\omega t - \delta/2)}]$$
$$= 2E_0 e^{i\delta/2} \cos(\omega t - \delta/2). \tag{5.16}$$

Hence

$$\begin{aligned} E_x &= 2E_0 \cos(\delta/2) \cos(\omega t - \delta/2), \\ E_y &= 2E_0 \sin(\delta/2) \cos(\omega t - \delta/2). \end{aligned} \tag{5.17}$$

It follows from (5.17) that $E_y/E_x = \tan(\delta/2)$, i.e. the resultant wave is linearly polarized, the line of oscillations of vector E forming an angle $\delta/2$ with the X-axis.

SEC. 6. AVERAGING

It is shown that the result of averaging depends on the averaging interval. Computation rules for complex representation of the waves are described.

AVERAGING OPERATION. The quantities used in physical theories pertain to a point in space and time. The quantities measured in a physical experiment, however, represent an average over a certain volume and a certain interval of time. For example, the instantaneous value of a quantity is its limiting value as the interval of time over which the averaging is carried out tends to zero. However, this limiting value cannot be attained in an experiment. The result of measurement is an averaging over a small interval of time. All that has been said about the time averaging applies to the volume averaging as well.

The mean value of a physical quantity $f(x,y,z,t)$ is defined by the formula

$$\boxed{\langle f(x,y,z,t)\rangle_{\Delta V,\Delta\tau} = \frac{1}{\Delta V\Delta\tau}\int\limits_{\Delta V,\Delta\tau} f(x,y,z,t)\,\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t,} \tag{6.1}$$

where ΔV and $\Delta\tau$ are the volume and time interval over which the averaging is carried out.

The result of averaging depends on the averaging domain. Range of variation in f which is smaller than the averaging domain is not reflected in the averaged quantities. For example, the time variation of the electric field strength in a harmonic oscillation cannot be measured if the averaging is carried out over the period of oscillations.

If at any instant of time during the averaging interval $\Delta\tau$, the quantity f varies insignificantly over the averaging domain ΔV and this variation can be neglected, the quantity f in (6.1) can be treated as constant while integrating with respect to x, y, z, and averaging over (x,y,z) can be carried out:

$$\langle f(x,y,z,t)\rangle_{\Delta V,\Delta\tau} = \frac{1}{\Delta\tau}\int\limits_{\Delta\tau}\mathrm{d}t\frac{1}{\Delta V}\int\limits_{\Delta V} f(x,y,z,t)\,\mathrm{d}x\mathrm{d}y\mathrm{d}z$$

$$= \frac{1}{\Delta\tau}\int\limits_{\Delta\tau} f(x,y,z,t)\,\mathrm{d}t, \tag{6.2}$$

where $\frac{1}{\Delta V}\int\limits_{\Delta V}\mathrm{d}x\mathrm{d}y\mathrm{d}z = 1$. Thus, all averaging operations in this case can be reduced to time averaging:

$$\boxed{\langle f(t)\rangle_{\Delta\tau} = \frac{1}{\Delta\tau}\int\limits_{\Delta\tau} f\,\mathrm{d}t,} \tag{6.3}$$

where the spatial variables are omitted for the sake of simplicity of notation, since the averaging operation is independent of these variables.

AVERAGING OF HARMONIC FUNCTIONS. By definition,

$$\langle \mathrm{e}^{i\omega t}\rangle_{\Delta\tau} = \frac{1}{\Delta\tau}\int\limits_{t-\Delta\tau/2}^{t+\Delta\tau/2}\mathrm{e}^{i\omega t}\,\mathrm{d}t = \frac{1}{i\omega\Delta\tau}\mathrm{e}^{i\omega t}\Big|_{t-\Delta\tau/2}^{t+\Delta\tau/2}$$

$$= \frac{\sin(\omega\Delta\tau/2)}{(\omega\Delta\tau/2)}\mathrm{e}^{i\omega t}, \tag{6.4}$$

where $\mathrm{e}^{i\alpha} - \mathrm{e}^{-i\alpha} = 2i\sin\alpha$. Separating the real and imaginary parts in (6.4), we obtain

$$\langle \cos \omega t \rangle_{\Delta\tau} = \frac{\sin(\omega\Delta\tau/2)}{(\omega\Delta\tau/2)} \cos \omega t, \tag{6.5}$$

$$\langle \sin \omega t \rangle_{\Delta\tau} = \frac{\sin(\omega\Delta\tau/2)}{(\omega\Delta\tau/2)} \sin \omega t. \tag{6.6}$$

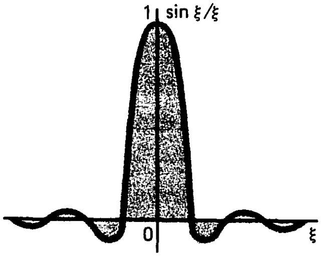


Fig. 20. Amplitude curve for an averaged harmonic function.

The result of averaging of a harmonic function is a harmonic function with the same frequency, but with an amplitude multiplied by

$$A(\xi) = \sin \xi/\xi \quad (\xi = \omega\Delta\tau/2) \tag{6.7}$$

(Fig. 20). It can be seen that the amplitude of the averaged harmonic function decreases rapidly with increasing $\Delta\tau$, and becomes equal to zero for $\xi = \omega\Delta\tau/2 = \pi$, $\Delta\tau = 2\pi/\omega = T$ (T is the period of harmonic oscillations). In particular, in the optical range of electromagnetic oscillations ($T \simeq 10^{-15}$ s) there are no instruments available at present for detecting electric field strength or magnetic induction over a period of time smaller than 10^{-15} s. Hence it is impossible to experimentally determine the variation of the electric field strength and the magnetic induction of a light wave. In optics, only the quantities whose values have been averaged over many oscillation periods are used.

AVERAGING OF SQUARES OF HARMONIC FUNCTIONS. In this case, we have

$$\langle \cos^2 \omega t \rangle_{\Delta\tau} = \langle \frac{1}{2}(1 + \cos 2\omega t) \rangle_{\Delta\tau}$$

$$= \frac{1}{2}\left[1 + \frac{\sin(\omega\Delta\tau)}{(\omega\Delta\tau)} \cos 2\omega t\right]. \tag{6.8}$$

$$\langle \sin^2 \omega t \rangle_{\Delta\tau} = \langle \frac{1}{2}(1 - \cos 2\omega t) \rangle_{\Delta\tau}$$

$$= \frac{1}{2}\left[1 - \frac{\sin(\omega\Delta\tau)}{(\omega\Delta\tau)} \cos 2\omega t\right], \tag{6.9}$$

where we have used double-angle formulas as well as relations (6.5) and (6.6). The plot of the mean square of a harmonic function is shown in Fig. 21. It can be seen that the mean value oscillates about 1/2 with an amplitude determined by the characteristic factor $A(\xi)$. With increasing interval of time over which averaging is carried out, the oscillation amplitude decreases and the mean square value of the harmonic function tends to the constant value 1/2. This situation is always encountered in the optical range.

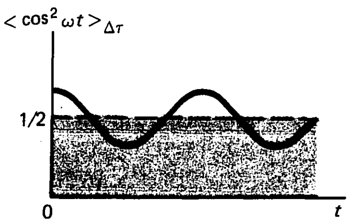


Fig. 21. Oscillation curve for the mean square of a harmonic function.

LINEARITY OF THE AVERAGING OPERATION. It follows from definition (6.1) that

$$\langle \mu_1 f_1 + \mu_2 f_2 \rangle = \langle \mu_1 f_1 \rangle + \langle \mu_2 f_2 \rangle = \mu_1 \langle f_1 \rangle + \mu_2 \langle f \rangle, \quad (6.10)$$

where μ_1 and μ_2 are constants. This equation indicates that the averaging is a linear operation.

CALCULATIONS INVOLVING COMPLEX SCALARS. For simplifying the calculations, oscillations and waves are often expressed in the complex form. The calculations of time averages have to be made by using the products of real or imaginary parts of complex quantities. However, the product of real parts of two complex quantities is not equal to the real part of their product, i. e.

$$\operatorname{Re} A \operatorname{Re} B \neq \operatorname{Re}(AB). \quad (6.11)$$

Hence the determination of the mean value of the product of real parts of two complex quantities cannot be reduced to the determination of the mean value of the real part of their product. However, the computation is considerably simplified in spite of this, since the time dependence of the quantities under consideration has the form

$$\begin{aligned} A(x,y,z,t) &= A_0(x,y,z)\,e^{-i\omega t}, \\ B(x,y,z,t) &= B_0(x,y,z)\,e^{-i\omega t}. \end{aligned} \quad (6.12)$$

Taking into account the equalities

$$\operatorname{Re} A = (1/2)(A + A^*), \quad \operatorname{Re} B = (1/2)(B + B^*), \quad (6.13)$$

we obtain

$$\begin{aligned} &\operatorname{Re} A \operatorname{Re} B = \operatorname{Re}(A_0 e^{-i\omega t}) \operatorname{Re}(B_0 e^{-i\omega t}) \\ &= (1/2)\operatorname{Re}(A_0 B_0^* + A_0 B_0\, e^{-2i\omega t}), \end{aligned} \quad (6.14)$$

where we have taken into consideration the relation

$$\operatorname{Re}(AB^*) = \operatorname{Re}(A^* B). \quad (6.15)$$

The time averaging of (6.14) gives

$$\begin{aligned} &\langle \operatorname{Re} A \operatorname{Re} B \rangle = (1/2)\operatorname{Re}(A_0 B_0^*) = (1/2)\operatorname{Re}(AB^*) \\ &= (1/2)\operatorname{Re}(A^* B), \end{aligned} \quad (6.16)$$

where A and B have the form given by (6.12).

CALCULATIONS INVOLVING COMPLEX VECTOR QUANTITIES. If the field vectors of a wave are represented in the complex form, i. e.

$$\mathbf{E} = \mathbf{E}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{r})}, \quad (6.17)$$

$$\mathbf{H} = \mathbf{H}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{r})}, \quad (6.18)$$

the energy flux density is defined by the formula

!

The result of averaging depends on the averaging interval. The mean values cannot be used to study the variations taking place over intervals smaller than the averaging interval.

?

Prove that the mean value of a derivative is equal to the derivative of the mean values. Why is the mean value of the product of real parts of two complex quantities not equal to the real part of the mean value of their product?

$$\boxed{\mathrm{S} = \mathrm{Re}\mathrm{E} \times \mathrm{Re}\mathrm{H}.} \tag{6.19}$$

We represent the time-independent parts of the vectors in the form

$$\mathrm{E}_0 e^{i\mathbf{k}\cdot\mathbf{r}} = \mathrm{E}_R + i\mathrm{E}_I,$$
$$\mathrm{H}_0 e^{i\mathbf{k}\cdot\mathbf{r}} = \mathrm{H}_R + i\mathrm{H}_I, \tag{6.20}$$

where the subscripts R and I stand for the real and imaginary parts of the quantities on the left-hand side of the equality. We have

$$\mathrm{E} = (\mathrm{E}_R + i\mathrm{E}_I)\,e^{-i\omega t}, \quad \mathrm{H} = (\mathrm{H}_R + i\mathrm{H}_I)\,e^{-i\omega t} \tag{6.21}$$

and hence

$$\mathrm{Re}\mathrm{E} = \mathrm{Re}\,[(\mathrm{E}_R + i\mathrm{E}_I)(\cos\omega t - i\sin\omega t)]$$
$$= \mathrm{E}_R\cos\omega t + \mathrm{E}_I\sin\omega t, \tag{6.22a}$$
$$\mathrm{Re}\mathrm{H} = \mathrm{H}_R\cos\omega t + \mathrm{H}_I\sin\omega t. \tag{6.22b}$$

Substituting these expressions into (6.19), we obtain

$$\mathrm{S} = (\mathrm{E}_R \times \mathrm{H}_R)\cos^2\omega t \times (\mathrm{E}_I \times \mathrm{H}_I)\sin^2\omega t$$
$$+ [(\mathrm{E}_R \times \mathrm{H}_I) + (\mathrm{E}_I \times \mathrm{H}_R)]\sin\omega t\cos\omega t. \tag{6.23}$$

During time averaging, the third term on the right-hand side of this equality vanishes, while the mean square values of the sine and cosine are equal to 1/2. Hence

$$\langle \mathrm{S} \rangle = (1/2)(\mathrm{E}_R \times \mathrm{H}_R) + (1/2)(\mathrm{E}_I \times \mathrm{H}_I). \tag{6.24}$$

This relation can be expressed in a simpler form by using complex conjugates. It follows from (6.21) that

$$\mathrm{H}^* = (\mathrm{H}_R - i\mathrm{H}_I)\,e^{i\omega t}. \tag{6.25}$$

Let us calculate the vector product:

$$\mathrm{E} \times \mathrm{H}^* = (\mathrm{E}_R + i\mathrm{E}_I)\,e^{-i\omega t} \times (\mathrm{H}_R - i\mathrm{H}_I)\,e^{i\omega t}$$
$$= (\mathrm{E}_R \times \mathrm{H}_R) + \mathrm{E}_I \times \mathrm{H}_I + i[(\mathrm{E}_I \times \mathrm{H}_R) - (\mathrm{E}_R \times \mathrm{H}_I)]. \tag{6.26}$$

Comparing (6.26) and (6.24), we get

$$\boxed{\langle \mathrm{S} \rangle = (1/2)\,\mathrm{Re}\,(\mathrm{E} \times \mathrm{H}^*).} \tag{6.27}$$

Similarly, we can calculate the mean volume density of the energy of electric and magnetic fields of a wave:

$$\boxed{\langle w_{\mathrm{el}} \rangle = (\varepsilon/4)\,\mathrm{E}\cdot\mathrm{E}^*,} \tag{6.28}$$

$$\boxed{\langle w_{\mathrm{m}} \rangle = (\mu/4)\,\mathrm{H}\cdot\mathrm{H}^* = [1/(4\mu)]\,\mathrm{B}\cdot\mathrm{B}^*.} \tag{6.29}$$

SEC. 7. PHOTOMETRIC CONCEPTS AND QUANTITIES

The method of converting radiant quantities into photometric quantities and vice versa is described.

RADIANT AND PHOTOMETRIC QUANTITIES. In the optical range, physical instruments and the human eye register the values of measured quantities averaged over a large number of oscillation periods. The mean values of the electric field strength and the magnetic induction are equal to zero and cannot be determined at any instant of time. The simplest quantities that can be registered are those which depend on the squares of strength, i. e. radiant quantities like the volume density of the radiant energy, radiant flux density, emissive power, and other quantities derived from them. These quantities are measured with the help of physical instruments.

It should be noted that the electric field of a wave does not act in the same way as the magnetic field. For example, it is well known that the darkening of a photographic plate by light is caused by the electric field of a wave. However, considering that the volume densities of the energy of the electric and magnetic fields are equal, we can always characterize the action of a wave in terms of radiant quantities.

The human eye also registers averaged values, i. e. reacts to the radiant properties of the radiation in the visible range. In other words, the human eye responds to the radiant characteristics of light. However, the perception caused by light depends not only on its radiant characteristics, but also on other circumstances, mainly on the wavelength of the light. For example, the eye is most sensitive to the green light with a wavelength of 555 nm. The sensitivity of the eye to light falls to zero at the boundaries of the visible range. For example, the emissive power of radiation of wavelength 760 nm must be increased by a factor of about 20 000 so that it creates the same perception of brightness to the human eye as radiation of wavelength 555 nm.

In many cases, it is not the radiant characteristics of light themselves that are of interest, but rather the subjective perceptions associated with them. For example, it is necessary to calculate the optimum illumination of the writing desk. This cannot be done with the help of radiant characteristics since radiation of the same power directed towards the table will cause different perceptions of illuminance for different spectral composition of light. This problem can be solved by using some quantities other than radiant quantities, namely photometric quantities. Radiant and photometric quantities are interrelated.

RADIANT QUANTITIES. The definitions of the radiant quantities are based on the emissive power. If an energy $\mathrm{d}W$ is emitted in the form of radiation over a time $\mathrm{d}t$, the emissive power will be given by

$$P = \mathrm{d}W/\mathrm{d}t. \tag{7.1a}$$

The emissive power is distributed over all possible wavelengths.

Spectral emissive power is the quantity

$$P_\lambda = \mathrm{d}P/\mathrm{d}\lambda, \tag{7.1b}$$

where $\mathrm{d}P$ is the power in the wavelength interval $(\lambda, \lambda + \mathrm{d}\lambda)$. Obviously

$$\mathrm{d}P = P_\lambda \mathrm{d}\lambda. \tag{7.1c}$$

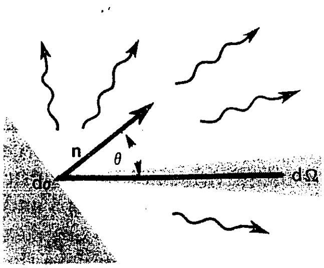


Fig. 22. To the definition of quantities characterizing the radiation emitted by a surface element.

The surfaces of material objects emit radiation. An elementary radiator is a surface element of a body with an area $\mathrm{d}\sigma$ (Fig. 22). In this section, $\mathrm{d}\sigma$ denotes the area of the surface element [since the letter S is used for the energy flux density (3.4)] and $\mathrm{d}P$ is the emissive power of an elementary source.

RADIANT INTENSITY. The radiant intensity $\mathrm{d}I$ of an elementary source is the ratio of the emissive power $\mathrm{d}P$ in an element $\mathrm{d}\Omega$ of solid angle to $\mathrm{d}\Omega$:

$$\mathrm{d}I = \mathrm{d}P/\mathrm{d}\Omega. \tag{7.2a}$$

For the spectral density of radiation, this formula assumes the form

$$\mathrm{d}I_\lambda = \mathrm{d}P_\lambda/\mathrm{d}\Omega, \tag{7.2b}$$

where $\mathrm{d}I_\lambda = \mathrm{d}(\mathrm{d}I)/\mathrm{d}\lambda$ is the spectral radiance in the wavelength interval $(\lambda, \lambda + \Delta\lambda)$.

Obviously, $\mathrm{d}I$ depends in general on the direction of radiation, i. e. on the orientation of the solid angle element $\mathrm{d}\Omega$ relative to the elementary radiator. If the elementary radiator is a surface element of a body, then $\mathrm{d}I$ depends, among other things, on the angle θ between the normal n to the surface and the direction in which the solid angle element $\mathrm{d}\Omega$ is oriented (Fig. 22), as well as on the axial angle characterizing the rotation about the normal as the axis. The normal from the surface is in the direction of emission of the radiation. Moreover, it is obvious that $\mathrm{d}I$ is also proportional to the area $\mathrm{d}\sigma$ of the surface element.

The radiant intensity of a point source emitting radiation

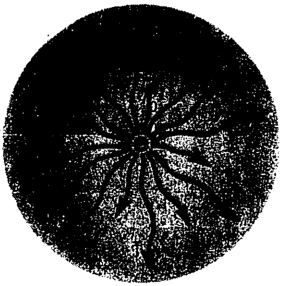


Fig. 23. A point source emitting radiation uniformly in all directions.

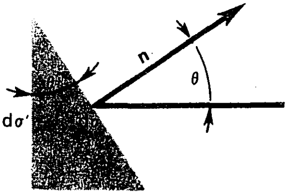


Fig. 24. To the definition of radiance.

uniformly in all directions (Fig. 23) is

$$\mathrm{d}I = I_0 = \text{const.} \tag{7.3}$$

Formula (7.2a) leads to the relation

$$P = \int_{\Omega=4\pi} \mathrm{d}P = \int_{\Omega=4\pi} I_0 \mathrm{d}\Omega = 4\pi I_0, \tag{7.4}$$

connecting the radiant intensity of the point source with its total emissive power.

RADIANCE. Radiation from a surface element dσ is emitted in all directions characterized by the angle θ between the normal n to the surface element and the direction of propagation of the radiation (Fig. 24). The projection of dσ onto the surface perpendicular to the direction of propagation of radiation is

$$\mathrm{d}\sigma' = \mathrm{d}\sigma \cos\theta. \tag{7.5}$$

The radiance of a surface at a point on the element dσ is the ratio of radiant intensity dI from this element to the area dσ':

$$L = \frac{\mathrm{d}I}{\mathrm{d}\sigma'} = \frac{\mathrm{d}I}{\mathrm{d}\sigma\cos\theta} = \frac{\mathrm{d}P}{\mathrm{d}\Omega\,\mathrm{d}\sigma\cos\theta}. \tag{7.6a}$$

The formula for spectral radiance has the form

$$L_\lambda = \frac{\mathrm{d}I_\lambda}{\mathrm{d}\sigma'} = \frac{\mathrm{d}I_\lambda}{\mathrm{d}\sigma\cos\theta} = \frac{\mathrm{d}P_\lambda}{\mathrm{d}\Omega\,\mathrm{d}\sigma\cos\theta}. \tag{7.6b}$$

Obviously, the radiance depends on the direction of radiation and is generally different at different points on the surface.

RADIANT EMITTANCE (EXITANCE). The radiant flux per unit area that emerges from a surface in all possible directions is called the radiant emittance:

$$M = \frac{\mathrm{d}P}{\mathrm{d}\sigma} = \int_{\Omega=2\pi} \frac{\mathrm{d}I\mathrm{d}\Omega}{\mathrm{d}\sigma} = \int_{\Omega=2\pi} L\cos\theta\,\mathrm{d}\Omega, \tag{7.7a}$$

where the integration is carried out over the solid angle 2π to cover all directions from the element dσ on the side of radiation emission.

The spectral radiant emittance is defined by the formula

$$M_\lambda = \frac{\mathrm{d}P_\lambda}{\mathrm{d}\sigma} = \int_{\Omega=2\pi} \frac{\mathrm{d}I_\lambda\mathrm{d}\Omega}{\mathrm{d}\sigma} = \int_{\Omega=2\pi} L_\lambda\cos\theta\,\mathrm{d}\theta. \tag{7.7b}$$

If the spectral radiance L_λ is independent of the direction (i.e. if $L_\lambda = \text{const}$), the integral in (7.7b) can be evaluated.

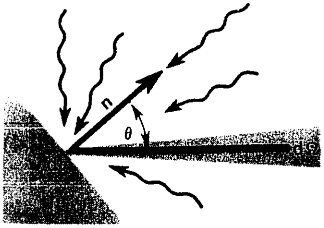


Fig. 25. To the calculation of radiant illuminance.

Directing the Z-axis of the spherical coordinate system along the normal to the surface element and denoting the axial angle by φ, we can write (7.7b) in the form

$$M_\lambda = L_\lambda \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi/2} \sin\theta\cos\theta\,\mathrm{d}\theta. \tag{7.8}$$

Integrating, we get

$$\underline{\overline{M_\lambda = \pi L_\lambda.}} \tag{7.9a}$$

This relation leads to the equality $\int_0^\infty M_\lambda \mathrm{d}\lambda = \pi \int_0^\infty L_\lambda \mathrm{d}\lambda$ which can be written in the following form by taking into account (7.6a), (7.6b), (7.7a) and (7.7b):

$$\underline{\overline{M = \pi L,}} \tag{7.9b}$$

where

$$\underline{\overline{M = \int_0^\infty M_\lambda \mathrm{d}\lambda, \quad L = \int_0^\infty L_\lambda \mathrm{d}\lambda}} \tag{7.9c}$$

are the radiant emittance and the radiance of the surface respectively.

RADIANT ILLUMINANCE. All the above-mentioned quantities characterize the emission process. Let us now consider the incidence of radiation on a surface element. This phenomenon is characterized by a quantity called the radiant illuminance, equal to the ratio of the radiation power dP incident on a surface element to the area dσ of this element:

$$\underline{\overline{E = \mathrm{d}P/\mathrm{d}\sigma.}} \tag{7.10a}$$

In calculations, the normal n to the surface is assumed to be directed towards the side of the surface dσ from which the radiation is falling (Fig. 25).

The spectral radiant illuminance is given by

$$\underline{\overline{E_\lambda = \mathrm{d}P_\lambda/\mathrm{d}\sigma.}} \tag{7.10b}$$

PHOTOMETRIC QUANTITIES are defined in the same way as the radiant quantities with the only difference that the luminous intensity is now taken as the fundamental quantity. The unit of luminous intensity, viz. a candela, is defined with the help of a blackbody radiator (taken as the basic standard) operating at the freezing point of platinum. This standard was

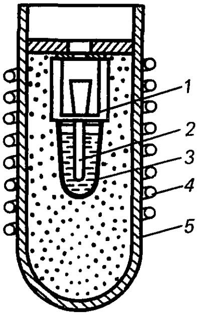


Fig. 26. The standard of luminous intensity.

adopted in 1967 at the XIII General Conference on Weights and Measures, and consists of a 40-mm-long ceramic tube *2* (Fig. 26) having a diameter up to 2 mm and closed at the lower end. This tube is placed in melting crucible *3* filled with pure platinum. For thermal insulation, the crucible is placed in vessel *5* containing thorium powder. Platinum is melted by induction currents excited by an alternating current flowing through coil *4*. Upon cooling, platinum freezes and its temperature is maintained at a constant value of 2045 K. The tube and the melting crucible are closed from above by lid *1* with a hole. The radiation emitted through this hole defines the unit of luminous intensity. This radiation is directed towards a surface which plays the role of a photometer. The luminous intensity of another source is determined by comparing the illuminance of this source with that of the standard.

A candela (cd) is the luminous intensity perpendicular to the surface of a blackbody radiator of area $(1/6) \times 10^{-5}$ m^2 at the freezing point of platinum under a pressure of 101 325 Pa.

A candela is the primary unit of luminous intensity. All other photometric quantities are derived from it. We shall denote these quantities by the same letters as the radiant quantities, adding the subscript V to these symbols. In most cases, photometric quantities are termed in the same way as the radiant quantities, except that the word “radiant” is replaced by “luminous” or its appropriate derivatives. Each photometric quantity has a radiant quantity corresponding to it, both having identical properties. The luminous intensity is denoted by $\mathrm{d}I_V$ and corresponds to the radiant intensity $\mathrm{d}I$ [see (7.2)].

LUMINOUS FLUX. Luminous flux is the product of the luminous intensity $\mathrm{d}I_V$ of a source by a solid angle $\mathrm{d}\Omega$ over which the light is emitted:

$$\boxed{\mathrm{d}\Phi_V = \mathrm{d}I_V \mathrm{d}\Omega.} \tag{7.11}$$

The luminous flux is analogous to the emissive power in the system of radiant quantities [see (7.2)], but is usually denoted by Φ_V instead of P_V. It can be seen from (7.11) that if a point source of luminous intensity I_{0V} emits in all directions, its total radiant flux will be $\Phi_V = 4\pi I_{0V}$ [see (7.4)].The spectral density of luminous flux is defined by the formula

$$\boxed{\mathrm{d}\Phi_{V\lambda} = \mathrm{d}I_{V\lambda} \mathrm{d}\Omega.} \tag{7.12}$$

LUMINANCE. This term is introduced in the same way as radiance (7.6):

$$L_V = \frac{\mathrm{d}I_V}{\mathrm{d}\sigma'} = \frac{\mathrm{d}I_V}{\mathrm{d}\sigma\cos\theta} = \frac{\mathrm{d}\Phi_V}{\mathrm{d}\Omega\mathrm{d}\sigma\cos\theta}, \tag{7.13a}$$

$$L_{V\lambda} = \frac{\mathrm{d}\Phi_{V\lambda}}{\mathrm{d}\Omega\mathrm{d}\sigma\cos\theta}. \tag{7.13b}$$

Equation (7.13a) can be rewritten in the form

$$\mathrm{d}I_V = L_V\mathrm{d}\sigma\cos\theta. \tag{7.14}$$

The dependence of L_V on the angle θ is governed by the properties of the surface. If L_V is independent of angles, then $L_V\mathrm{d}\sigma = (\mathrm{d}I)_{V0} = \text{const}$ and (7.14) assumes the form

$$\mathrm{d}I_V(\theta) = (\mathrm{d}I)_{V0}\cos\theta, \tag{7.15}$$

where $\mathrm{d}I_{V0}(\theta)$ is the luminous intensity in the direction of the angle θ and $(\mathrm{d}I)_{V0}$ is the luminous intensity along the normal to the surface. Dependence (7.15) is called Lambert's law. The surfaces for which the emitted radiation is characterized by the condition $L_V = \text{const}$ are called Lambert surfaces. The radiation emitted from such surfaces is diffused, and hence these surfaces are also called diffusively radiating surfaces.

As an example, let us consider the values of luminances of some sources: moonless night sky $\sim 10^{-4}$ cd/m^2, full moon $\sim 10^3$ cd/m^2, cloudless clear day sky $\sim 10^4$ cd/m^2, incandescent lamp filaments $\sim 10^6$ cd/m^2, the Sun $\sim 10^9$ cd/m^2.

LUMINOUS EMITTANCE (exitance). In analogy with (7.7), we have

$$M_V = \frac{\mathrm{d}\Phi_V}{\mathrm{d}\sigma} = \int\limits_{\Omega=2\pi} L_V\cos\theta\mathrm{d}\Omega,$$

$$M_{V\lambda} = \int\limits_{\Omega=2\pi} L_{V\lambda}\cos\theta\mathrm{d}\Omega. \tag{7.16}$$

ILLUMINANCE. In accordance with (7.10), the illuminance and its spectral density are defined by the relations

$$E_V = \frac{\mathrm{d}\Phi_V}{\mathrm{d}\sigma}, \tag{7.17}$$

$$E_{V\lambda} = \frac{\mathrm{d}\Phi_{V\lambda}}{\mathrm{d}\sigma}. \tag{7.18}$$

LIGHT EXPOSURE. This quantity is defined by the formula

$$dH_V = \Phi_V \, \mathrm{d}t, \tag{7.19}$$

where $\mathrm{d}t$ is the interval of time over which the luminous flux Φ_V is emitted or received.

RELATIONS BETWEEN RADIANT AND LUMINOUS CHARACTERISTICS OF RADIATION. The same emissive power in different wavelength regions may cause quite different visual perceptions concerning the brightness of the source. If the wavelengths lie beyond the visible part of the spectrum, the source remains invisible irrespective of its intensity. For a fixed power, the source is the brightest at a wavelength of 555 nm. If we are considering the illumination of a room, we cannot confine ourselves just to the energy aspect of the problem. We must also take into account the spectral composition of the light source as well as the relation between the real emissive power in different wavelength regions and the subjective visual perception of the emissive power of the source. In the first place, we must estimate the ratio of the spectral density $\Phi_{V\lambda}$ of the luminous flux and the spectral emissive power P_λ for identical visual perceptions at different wavelengths. Quantitatively, this relation is characterized by the ratio

$$V(\lambda) = \Phi_{V\lambda}/P_\lambda, \tag{7.20}$$

called the spectral luminous efficacy.The luminous flux Φ_V is estimated from the visual perception of the emissive power in lumens, while P is expressed in watts. The spectral luminous efficacy depends on the wavelength and is equal to zero for wavelengths lying beyond the visible range. In the visible range, its value varies slightly for various individuals. Measurements in which a large number of people with normal vision were involved led to a certain mean statistical curve $V(\lambda)$ which is accepted as a standard.

The quantity $V(\lambda)$ attains its peak value at the wavelength 555 nm. Instead of $V(\lambda)$, it is convenient to use the dimensionless quantity called the relative spectral luminous efficacy.This quantity is defined as

$$K(\lambda) = V(\lambda)/V(555\ \mathrm{nm}). \tag{7.21}$$

The K vs. λ curve is shown in Fig. 27. It is most "economical" to use light with wavelength 555 nm for illumination. This, however, is not acceptable since all the objects surrounding us would then appear green with different

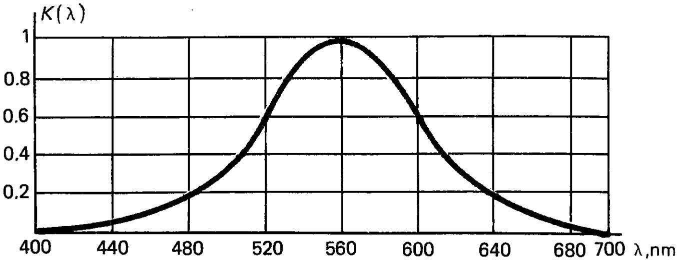


Fig. 27. Curve showing relative luminous efficacy.

intensities. The illumination must be accomplished by radiation with a spectral composition corresponding to the "mean sunlight" which differs from the spectral composition of solar radiation (see Sec. 1). Taking the economical efficiency of a source with wavelength 555 nm used for illumination as unity, we find that the economical efficiency of the source generating mean sunlight is 0.35. The corresponding values for the solar radiation, incandescent lamp and fluorescent lamp are 0.14, 0.02 and 0.06 respectively. If the spectral emissive power P_λ is known, the total emissive power is given by

$$P = \int_0^\infty P_\lambda \mathrm{d}\lambda. \tag{7.22a}$$

According to (7.20) and (7.21), the spectral density of the luminous flux is defined by

$$\Phi_{V\lambda} = V(555\ \mathrm{nm})\, K(\lambda)\, P_\lambda, \tag{7.22b}$$

while the luminous flux is expressed in the form

$$\Phi_V = \int_{\lambda_1}^{\lambda_2} \Phi_{V\lambda} \mathrm{d}\lambda = V(555\ \mathrm{nm}) \int_{\lambda_1}^{\lambda_2} K(\lambda)\, P_\lambda \mathrm{d}\lambda. \tag{7.23}$$

These formulas connect the values of radiant and photometric quantities. Outside the visible range, the factor $K(\lambda)$ is equal to zero. Let us apply formula (7.23) to the radiation from the standard which was used for the definition of candela. Since the standard emits radiation as blackbody at a temperature of 2045 K, we can calculate the right-hand side of (7.23) with the help of Planck's formula for this temperature, while the left-hand side can be determined with the help of the definition of the standard. Combining formulas (7.2), (7.11) and (7.23) we can go over to the luminous intensity:

$$\mathrm{d}I_V = V(555\ \mathrm{nm}) \int_{\lambda_1}^{\lambda_2} K(\lambda)\, \mathrm{d}I_\lambda \mathrm{d}\lambda. \tag{7.24}$$

The standard is defined in terms of the luminous intensity in a direction perpendicular to the emitting surface. The surface of a blackbody emits radiation like a Lambert surface. Hence we obtain the following expression for the luminous intensity from an area $d\sigma$ over an angle 2π, per unit surface area:

$$\frac{dI_V}{d\sigma} = \frac{2\pi}{(1/6)\,10^{-5}}\,\frac{\text{cd}}{\text{m}^2} = 2\pi\cdot 6\cdot 10^5\,\frac{\text{cd}\cdot\text{sr}}{\text{m}^2\cdot\text{sr}}$$
$$= 2\pi\cdot 6\cdot 10^5\ \text{lm}/(\text{m}^2\cdot\text{sr}). \tag{7.25}$$

The radiant intensity from the surface $d\sigma$, per unit area of this surface, is equal to

$$\frac{dI}{d\sigma} = \frac{M}{2\pi}, \tag{7.26}$$

since by definition the radiant emittance is the total radiant flux emitted in the half-space and referred to the area of the surface while the radiant intensity is the total radiant flux emitted per unit solid angle. Taking into consideration (7.25) and (7.26), we can write Eq. (7.24) in the form

$$2\pi\cdot 6\cdot 10^5\ \text{lm}/(\text{m}^2\cdot\text{sr}) = V(555\ \text{nm}) \int_{\lambda_1}^{\lambda_2} K(\lambda)\frac{c}{8\pi} w_\lambda d\lambda, \tag{7.27}$$

where we have used the relation (50.24) for connecting M_λ with the spectral density w_λ. The expression (50.28) can be written in the form

$$w_\lambda = \frac{16\pi^2 c\hbar}{\lambda^5}\,\frac{1}{\exp[2\pi\hbar c/(kT\lambda)] - 1}. \tag{7.28}$$

We substitute this expression into formula (7.27) with $T = 2045$ K. After this substitution, (7.27) can be integrated since $K(\lambda)$ is known. Calculations lead to the following result:

$$V(555\ \text{nm}) = 1/M = 683\ \text{lm/W}, \tag{7.29}$$

$$M(555\ \text{nm}) = 1.46 \times 10^{-3}\ \text{W/lm}. \tag{7.30}$$

The constants $V(555$ nm$)$ and $M(555$ nm$)$ are respectively called the photometric equivalent of radiation and the radiant equivalent of light.

Example 7.1. A surface element $d\sigma_0$ of a sphere of radius R emits light in accordance with Lambert's law. Find the illuminance of different points on the hollow surface (Fig. 28).

Let us calculate the illuminance of the surface element $d\sigma_1$. We draw the normals to the surface elements $d\sigma_0$ and $d\sigma_1$ towards the centre of the hollow sphere and denote by θ_0 and θ_1 the angles formed by these normals with the line joining the surface elements. The luminance of the surface element $d\sigma_0$ is denoted by L_V. The luminous intensity from the surface element is given by the formula

$$\mathrm{d}I_V = L_V \mathrm{d}\sigma_0 \cos\theta_0, \tag{7.31}$$

L_V being independent of the angle θ_0.

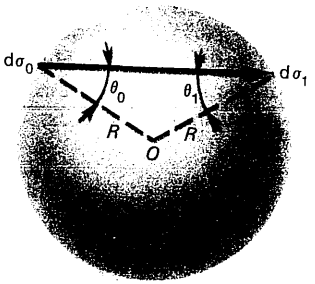


Fig. 28. To the calculation of illuminance of a hollow surface.

The illuminance of the element $\mathrm{d}\sigma_1$ is obtained on the basis of (7.17) and (7.11):

$$E_V = \mathrm{d}I_V \cos\theta_1 / r^2, \tag{7.32}$$

where r is the distance between the elements $\mathrm{d}\sigma_0$ and $\mathrm{d}\sigma_1$. The line joining the elements $\mathrm{d}\sigma_0$ and $\mathrm{d}\sigma_1$ is a chord of the circle at the section of the hollow sphere by the plane passing through the centre of the sphere and the elements $\mathrm{d}\sigma_0$ and $\mathrm{d}\sigma_1$. It can be seen that $\theta_0 = \theta_1$ and $r = 2R\cos\theta_0$. In view of this, Eq. (7.32) assumes the form

$$E_V = \frac{I_V \cos\theta_0}{4R^2\cos^2\theta_0} = \frac{L_V \mathrm{d}\sigma_0 \cos^2\theta_0}{4R^2\cos^2\theta_0} = \frac{L_V \mathrm{d}\sigma_0}{4R^2}, \tag{7.33}$$

where we have used relation (7.14). Thus, the radiant illuminance of all points on the surface of a hollow sphere is the same.

Example 7.2. A blackbody emits radiation from a plane circular surface of area 3 mm at the temperature of freezing platinum at a pressure of 101 325 Pa. Find the illuminance of another surface element forming an angle of 45° with the normal to the emitting surface and situated at a distance of 0.75 m from it. The angle between the normal to the surface receiving the radiation and the line joining the emitting and receiving surfaces is 30°.

To a fairly high degree of accuracy, the radiation emitted by a blackbody at the freezing point of platinum is equivalent to the radiation emitted by the standard used for defining a candela. The area of the emitting surface is 2.83×10^{-5} m^2. Hence the light of intensity $I_V(0°) = 2.83 \times 10^{-5} \times 6 \times 10^5$ cd $= 17$ cd is emitted along the normal to the surface. According to Lambert's law, the luminous intensity at an angle of 45° to the normal is $I_V(45°) = I_V(0°)\cos 45° = 12$ cd. The illuminance of a surface at a distance r from the emitter and inclined at an angle of 30° to the luminous flux is

$$E_V = \frac{I_V(45°)\cos 30°}{r^2} = \frac{12\cos 30°}{(0.75)^2}\,\frac{\mathrm{cd}}{\mathrm{m}^2} = 18.5\ \ \mathrm{lx}. \tag{7.34}$$

Example 7.3. Find the luminance of the solar surface if it is known that the direct incidence of solar radiation on a clear day creates on a surface normal to the direction to the Sun an

?

Is it possible to indicate the illuminance of a writing-table if the source of illumination has a very high radiant emittance?
How can the luminous emittance be increased without changing the emissive power of the source?
If the luminance of a surface element is independent of the direction (i.e. if we consider a Lambert surface), does it mean that the luminous flux from this surface is the same in all directions?

illuminance of 7×10^4 lx. The absorption and scattering of the solar radiation in the atmosphere can be neglected.

Since the Sun emits radiation like a blackbody, its luminance $L_V = \text{const}$ and hence there exists a relation similar to (7.9) between luminance and luminous emittance. This leads to the luminance of the solar surface

$$L_V = \frac{M_V}{\pi} = \frac{1}{\pi}\left(\frac{\mathrm{d}\Phi_V}{\mathrm{d}\sigma}\right)_0, \qquad (7.35)$$

where we have taken into consideration formula (7.16). Since the luminous flux from the Sun decreases in inverse proportion to the square of the distance and the solar radiation is isotropic, we obtain the following expression from (7.17) for the illuminance of a surface normal to the luminous flux:

$$E_V = \frac{\mathrm{d}\Phi_V}{\mathrm{d}\sigma} = \left(\frac{\mathrm{d}\Phi_V}{\mathrm{d}\sigma}\right)_0 \left(\frac{r}{R}\right)^2, \qquad (7.36)$$

where r is the radius of the Sun and R is the radius of the Earth's orbit. It is well known that the diameter of the Sun is seen from the Earth at an angle of 32′. Hence $r/R = 4.7 \times 10^{-3}$. Taking (7.36) into account, we obtain from (7.35):

$$L_V = \frac{1}{\pi} E_V \left(\frac{R}{r}\right)^2 = \frac{7 \times 10^4}{3.14(4.7 \times 10^{-3})^2} \frac{\text{cd}}{\text{m}^2} = 10^9\,\text{cd/m}^2. \qquad (7.37)$$

Example 7.4. Radiation of wavelength varying from $\lambda_1 = 555$ nm to $\lambda_2 = 600$ nm is incident on a plane surface at an angle of 45°. The spectral density of the radiant flux along the surface is $s_{\lambda 0} = 10^4$ W/m^3 at a wavelength of 555 nm on the wavelength scale and drops linearly to zero at a wavelength of 600 nm. Considering that the relative spectral luminous efficacy also drops linearly from $K_1 = 1$ to $K_2 = 0.5$ in this region of wavelengths, find the illuminance of the surface.

The dependence of the spectral density of the radiant flux on wavelength in the given interval is determined by the relation

$$s_\lambda = s_{\lambda 0} \frac{\lambda_2 - \lambda}{\lambda_2 - \lambda_1}, \qquad (7.38)$$

while the relative spectral luminous efficacy is given by

$$K(\lambda) = \frac{K_1(\lambda_2 - \lambda)}{\lambda_2 - \lambda_1} - \frac{K_2(\lambda_1 - \lambda)}{\lambda_2 - \lambda_1}.$$

The emissive power in the wavelength interval $\mathrm{d}\lambda$ incident on the surface $\mathrm{d}\sigma'$ is equal to

!

Radiant quantities are determined with the help of radiation power, while photometric quantities are determined from the luminous intensity. The radiation power corresponds to the luminous flux rather than the luminous intensity.
The conversion from radiant quantities to photometric ones is made with the help of spectral characteristics of radiation.

$$\mathrm{d}P = s_\lambda \mathrm{d}\lambda \mathrm{d}\sigma' = s_{\lambda 0}\frac{\lambda_2-\lambda}{\lambda_2-\lambda_1}\mathrm{d}\lambda \mathrm{d}\sigma', \tag{7.39}$$

where $\mathrm{d}\sigma' = \mathrm{d}\sigma\cos\theta$ is the area of the projection of the surface element $\mathrm{d}\sigma$ onto a plane perpendicular to the radiant flux incident on it. For the spectral density of the incident luminous flux, we obtain:

$$\mathrm{d}\Phi_{V\lambda} = \mathrm{d}P_\lambda V(555\ \mathrm{nm})K(\lambda), \tag{7.40}$$

where $V(555\ \mathrm{nm})$ is obtained from (7.29). Taking (7.39) into account, we find from (7.40):

$$E_{V\lambda} = \frac{\mathrm{d}\Phi_{V\lambda}}{\mathrm{d}\sigma} = s_\lambda V(555\ \mathrm{nm})K(\lambda)\cos\theta.$$

Hence the surface illuminance is given by

$$\begin{aligned}E_V &= \int_{\lambda_1}^{\lambda_2} E_\lambda \mathrm{d}\lambda = V(555\ \mathrm{nm})\int_{\lambda_1}^{\lambda_2} s_\lambda K(\lambda)\cos\theta\,\mathrm{d}\lambda \\ &= \frac{V(555\ \mathrm{nm})s_{\lambda 0}\cos\theta}{(\lambda_2-\lambda_1)^2}\int_{\lambda_1}^{\lambda_2}(\lambda_2-\lambda_1) \\ &\quad\times[K_1(\lambda_2-\lambda)-K_2(\lambda_1-\lambda)]\,\mathrm{d}\lambda \\ &= \frac{V(555\ \mathrm{nm})s_{\lambda 0}\cos\theta}{(\lambda_2-\lambda_1)^2}\left[\frac{1}{2}(K_1+K_2)(\lambda_2-\lambda_1)\right. \\ &\quad\left.\times(\lambda_1^2+\lambda_2^2-\lambda_1\lambda_2)+\frac{1}{3}(K_1-K_2)(\lambda_2^3-\lambda_1^3)\right].\end{aligned} \tag{7.41}$$

Substituting the numerical values of the quantities appearing in the formula and integrating, we obtain:

$$\begin{aligned}E_V &= \frac{683\times 10^4\times 0.701}{(45\times 10^{-9})^2}\left[0.5\times 1.5\times 45\times 10^{-9}(3.08\times 10^5\right. \\ &\quad + 3.6\times 10^5 - 3.3\times 10^5)\times 10^{-18} + \frac{1}{3}0.5(2.16\times 10^8 - 1.71 \\ &\quad\left.\times 10^8)\times 10^{-27}\right] = 54\,\mathrm{lx}.\end{aligned}$$

PROBLEMS

1.1. A sinusoidal voltage $U = U_0 \sin t$ is applied to the input of a rectifier. Find the Fourier series for the output voltage if the rectifier eliminates half-periods with negative values of the voltage.

1.2. A lamp without a reflector is hanging at a height of 0.5 m over the surface of a horizontal table. The total luminous flux of the lamp is 1000 lm and is distributed uniformly in all directions. What is the illuminance of the table along the vertical under the lamp?

1.3. Find the radiant emittance of the solar surface if the radiant flux density of the solar radiation on the Earth's orbit is 1400 W/m^2 and the diameter of the Sun is seen from the Earth at an angle of 32′.

1.4. The illuminance of the Earth's surface by the vertically incident solar radiation is 7×10^4 lx (the influence of the atmosphere is neglected). Find the luminous intensity of the Sun.

1.5. Using the data given in the Problem 1.3 for the solar radiation, determine the radiance of the solar surface.

1.6. The electric breakdown in air occurs at an electric field strength of about 3 MV/m. Find the energy flux density of plane electromagnetic waves for which a spark can be observed in air.

1.7. A point source of power 1 W emits photons of wavelength 0.5 μm uniformly in all directions. Find the photon flux density at a distance of 1 km from the source.

1.8. A blackbody at a temperature of 1500 K emits a luminous flux in a half-space. Find the illuminance at a distance of 1 m along the perpendicular to the surface if the radiation is incident normally on the surface.

1.9. Find the amplitude of the electric and magnetic field strengths of a wave having an energy flux density of 1 W/m^2.

1.10. Find the amplitude of the magnetic field strength of a wave propagating in water ($n = 1.33$). Also find the energy flux density of the waves if the electric field has an amplitude of 20 kV/m.

1.11. An elliptically polarized wave of energy flux density 20 W/m^2 propagates in water ($n = 1.33$). The eccentricity of the ellipse described by the tip of the field vector is 0.6. Find the amplitude of oscillations of the electric field vector components along the principal axes of the ellipse.

1.12. What is the light pressure when light is normally incident on a totally reflecting surface if the energy flux density of the waves is 1 W/m^2?

1.13. Find the electric and magnetic field strengths of a wave with wavelength 555 nm if it creates illuminance of 500 lx in the case of normal incidence.

ANSWERS

1.1. $$U_{\text{output}} = \frac{U_0}{\pi}\left(1 + \frac{\pi}{1 \times 2}\sin t - \frac{2}{1 \times 3}\cos 2t - \frac{2}{3 \times 5}\cos 4t - \frac{2}{5 \times 7} \times \cos 6t - \ldots\right).$$ 1.2. 321 lx. 1.3. 63.4 MW/m^2. 1.4. 1.6×10^{27} cd. 1.5. 2×10^7 W/(m$^2 \cdot$sr). 1.6. 1.2×10^{10} W/m^2. 1.7. 2×10^{11} photons/(m$^2 \cdot$s). 1.8. 17.3 lx. 1.9. 27.45 V/m; 0.073 A/m. 1.10. 70.6 A/m; 706 kW/m^2. 1.11. 78.9 V/m; 73.4 V/m. 1.12. 0.66×10^{-8} Pa. 1.13. 23.5 W/m; 7.6×10^{-2} A/m.

Chapter 2

Nonmonochromatic and Random Radiation

Basic idea:

Real radiation has a finite duration and propagates with randomly varying amplitude and phase. Its spectral composition is analyzed with the help of Fourier transforms.

SEC. 8. SPECTRAL COMPOSITION OF FUNCTIONS

The theory of Fourier integrals and series is used for analyzing the spectral composition of functions. The relation between the pulse duration and the spectral width is analyzed.

FOURIER SERIES IN REAL FORM. It is proved in mathematics that for quite general conditions that are normally satisfied in physical problems, a periodic function $f(t)$ with a period T, $[f(t+T)=f(t)]$ (Fig. 29) can be represented in the form of a Fourier series:

$$f(t)=\frac{a_0}{2}+\sum_{n=1}^{\infty}(a_n\cos n\,\omega t+b_n\sin n\,\omega t), \tag{8.1}$$

where

$$\omega=2\pi/T, \tag{8.2a}$$

$$a_n=\frac{2}{T}\int\limits_{-T/2}^{T/2}f(t)\cos n\,\omega t\,\mathrm{d}t, \tag{8.2b}$$

$$b_n=\frac{2}{T}\int\limits_{-T/2}^{T/2}f(t)\sin n\,\omega t\,\mathrm{d}t. \tag{8.2c}$$

A periodic position function $f(z)$ with a period L, $[f(z+L)=f(z)]$ (Fig. 30), has the same form as (8.1), but with the substitution $\omega \rightarrow k = 2\pi/L$, $T \rightarrow L$ and $t \rightarrow z$.

FOURIER SERIES IN COMPLEX FORM. Using the Euler formula $e^{i\alpha} = \cos\alpha + i\sin\alpha$, we can present the series (8.1) in the form

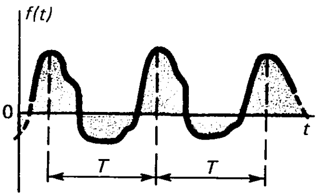


Fig. 29. A periodic time function.

$$f(t) = \sum_{n=-\infty}^{\infty} c\,e^{in\omega t}, \tag{8.3}$$

where

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,e^{-in\omega t}\,dt. \tag{8.4}$$

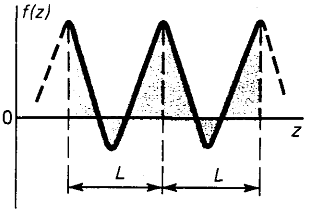


Fig. 30. A periodic position function.

The coefficients a_n, b_n and c_n are connected through the relations

$$\begin{gathered} c_0 = (1/2)\,a_0,\ c_n = (1/2)\,(a_n - ib_n)\ (n = 1, 2, 3, \ldots), \\ c_n = (1/2)(a_{-n} + ib_{-n})\ (n = -1, -2, -3, \ldots). \end{gathered} \tag{8.5}$$

FOURIER INTEGRAL IN REAL FORM. A nonperiodic function $f(t)$ (Fig. 31) cannot be presented in the form of a Fourier series. If the function $f(t)$ is piecewise continuous, has a finite number of extrema and is absolutely integrable in the interval $(-\infty, +\infty)$, it can be represented by a Fourier integral

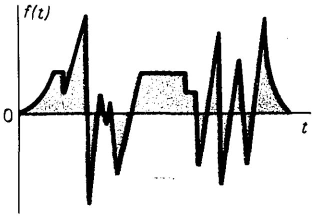


Fig. 31. A nonperiodic time function.

$$f(t) = \frac{1}{\pi}\int_0^{\infty} d\omega \int_{-\infty}^{\infty} f(\tau)\cos[\omega(t-\tau)]\,d\tau, \tag{8.6}$$

which is obtained from (8.1) for $T \rightarrow \infty$.

FOURIER INTEGRAL IN COMPLEX FORM. Relation (8.6) can be transformed with the help of Euler's formula:

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\,e^{i\omega t}\,d\omega, \tag{8.7}$$

where

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-i\omega t}\, \mathrm{d}t \tag{8.8}$$

is the Fourier transform of the function $f(t)$.

AMPLITUDE SPECTRUM AND PHASE SPECTRUM. The terms of the Fourier series (8.1) can be transformed as follows:

$$\begin{aligned} & a_n \cos n\,\omega t + b_n \sin n\,\omega t = \sqrt{a_n^2 + b_n^2} \\ & \times [(a_n/\sqrt{a_n^2 + b_n^2}) \cos n\,\omega t + (b_n/\sqrt{a_n^2 + b_n^2}) \sin n\,\omega t] \\ & = A_n \cos (n\,\omega t - \varphi_n), \end{aligned} \tag{8.9}$$

$$\begin{gathered} \cos \varphi_n = a_n/\sqrt{a_n^2 + b_n^2}, \ \sin \varphi_n = b_n/\sqrt{a_n^2 + b_n^2}, \\ A_n = \sqrt{a_n^2 + b_n^2}, \ \tan \varphi_n = b_n/a_n. \end{gathered} \tag{8.10}$$

As a result, formula (8.1) assumes the form

$$f(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos (n\,\omega t - \varphi_n), \tag{8.11}$$

where $A_0 = a_0$. The set of quantities A_n is called the amplitude spectrum of the function $f(t)$, while the set of quantities φ_n is called the phase spectrum of the function. By definition, the frequencies have positive values.

DETERMINATION OF THE AMPLITUDE- AND PHASE SPECTRA FROM COMPLEX FORM OF FOURIER SERIES. For a real function $f(t)$, the equality $c_n^* = c_{-n}$ is satisfied in (8.3). Transforming (8.3) into (8.1) and comparing the coefficients of $\cos n\omega t$ and $\sin n\omega t$, we obtain

$$a_0 = 2c_0, \ a_n = 2\mathrm{Re}\, c_n, \ b_n = -2\mathrm{Im}\, c_n, \tag{8.12}$$

where Re and Im represent the real and imaginary parts of a complex number. Combining (8.10) and (8.12), we obtain

$$A_0 = 2c_0, \ A_n = 2|c_n|, \ \tan \varphi_n = -\mathrm{Im}\, c_n/\mathrm{Re}\, c_n. \tag{8.13}$$

CONTINUOUS SPECTRUM. The Fourier transform (8.8) is called the complex spectrum, or simply spectrum, of the function $f(t)$. It completely defines the function $f(t)$ and is equivalent to the amplitude- and phase spectra. We have

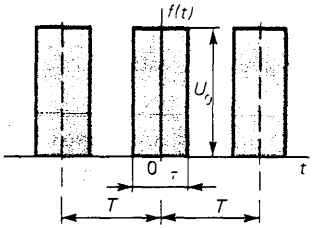


Fig. 32. An infinite sequence of rectangular pulses.

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)\,e^{i\omega t}d\omega=\frac{1}{2\pi}\int_{0}^{\infty}[F(\omega)\,e^{i\omega t}+F(-\omega)\,e^{-i\omega t}]\,d\omega. \tag{8.14}$$

For a real function $f(t)$, the relation $F^*(\omega)=F(-\omega)$ is satisfied, and hence we can write

$$F(\omega)\,e^{i\omega t}+F(-\omega)\,e^{-i\omega t}=F(\omega)\,e^{i\omega t}+[F(\omega)\,e^{i\omega t}]^*$$
$$=2\mathrm{Re}\,[F(\omega)\,e^{i\omega t}].$$

As a result, (8.14) assumes the form

$$f(t)=\frac{1}{\pi}\int_{0}^{\infty}[\mathrm{Re}\,F(\omega)\cos\omega t-\mathrm{Im}\,F(\omega)\sin\omega t]\,d\omega. \tag{8.15}$$

Transforming the integrand in (8.15) in accordance with formulas (8.9) and (8.10), we obtain

$$f(t)=\frac{1}{2\pi}\int_{0}^{\infty}A(\omega)\cos[\omega t-\varphi(\omega)]\,d\omega, \tag{8.16}$$

where

$$A(\omega)=2\,|F(\omega)|,\ \tan\varphi(\omega)=-\mathrm{Im}\,F(\omega)/\mathrm{Re}\,F(\omega). \tag{8.17}$$

The quantities $A(\omega)$ and $\varphi(\omega)$ represent the amplitude- and phase spectra of the function $f(t)$. The presence of the factor $1/(2\pi)$ in (8.16) indicates that $A(\omega)$ is the amplitude density referred to the frequencies $\nu=\omega/(2\pi)$ and not to the circular frequencies, since formula (8.16) can be presented in the form

$$f(t)=\int_{0}^{\infty}A(2\pi\nu)\cos[2\pi\nu t-\varphi(2\pi\nu)]\,d\nu. \tag{8.18}$$

Thus, periodic functions are characterized by discrete spectra while nonperiodic functions are described by continuous spectra.

SPECTRUM OF RECTANGULAR PULSES. Suppose that we have an infinite sequence of rectangular pulses of magnitude U_0 and duration τ. Let us suppose that these pulses are repeated after a time interval T (Fig. 32). Since the function is even, $b_n=0$, while the coefficients of the cosine are given by

$$a_n=\frac{2}{T}U_0\int_{-\tau/2}^{\tau/2}\cos n\omega t\,dt=2\frac{\tau}{T}U_0\,\frac{\sin(n\omega\tau/2)}{n\omega\tau/2}, \tag{8.19}$$

where $\omega = 2\pi/T$. Hence

$$A_n = 2\frac{\tau}{T}U_0\left|\frac{\sin(n\omega\tau/2)}{n\omega\tau/2}\right|, \quad \varphi_n = 0. \tag{8.20}$$

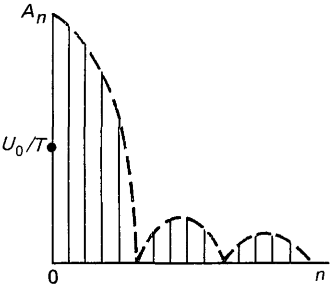


Fig. 33. Amplitude spectrum of an infinite sequence of rectangular pulses.

The amplitude spectrum is shown in Fig. 33. Since $n\omega\tau/2 = n\pi\tau/T$, the number of harmonics between zeros of the amplitude depends on the ratio τ/T: the shorter the pulse duration in comparison with the pulse-repetition period, the larger the number of harmonics ($\Delta n = T/\tau$).

It is worth noting that the Fourier series depends on the shape of the curve it represents, as well as on the position of the origin of coordinates relative to this curve. For example, if the origin of coordinates in Fig. 32 is shifted from the centre of a pulse to a point near its boundary, the periodic function will not have a definite evenness relative to this origin. Hence its Fourier series will be different from (8.32) since it will contain terms with $\cos n\omega t$ as well as those with $\sin n\omega t$.

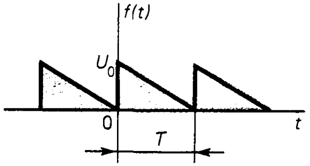


Fig. 34. An infinite sequence of sawtooth pulses.

SPECTRUM OF SAWTOOTH PULSES. Let us consider an infinite sequence of sawtooth pulses (Fig. 34).

We have

$$a_0 = \frac{2}{T}U_0\int_{-T/2}^{T/2} f(t)\,\mathrm{d}t = \frac{2}{T}U_0\int_0^T\left(1-\frac{t}{T}\right)\mathrm{d}t = U_0.$$

$$a_n = \frac{2}{T}U_0\int_0^T\left(1-\frac{t}{T}\right)\cos n\omega t\,\mathrm{d}t = 0, \tag{8.21}$$

$$b_n = \frac{2}{T}U_0\int_0^T\left(1-\frac{t}{T}\right)\sin n\omega t\,\mathrm{d}t = \frac{U_0}{\pi}\frac{1}{n}.$$

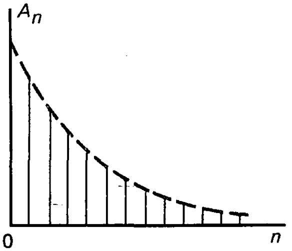


Fig. 35. Amplitude spectrum of an infinite sequence of sawtooth pulses.

The amplitude- and phase spectra are given by the formulas

$$A_n = U_0/(\pi n), \qquad \varphi_n = \pi/2. \tag{8.22}$$

The amplitude spectrum is shown in Fig. 35. It can be seen that the amplitudes decrease in inverse proportion to the number of the harmonic.

SPECTRUM OF A SINGLE RECTANGULAR PULSE. This pulse is shown in Fig. 36. We have

$$F(\omega) = U_0\int_{-\tau/2}^{\tau/2} \mathrm{e}^{-i\omega t}\,\mathrm{d}t = U_0\tau\frac{\sin(\omega\tau/2)}{\omega\tau/2}. \tag{8.23}$$

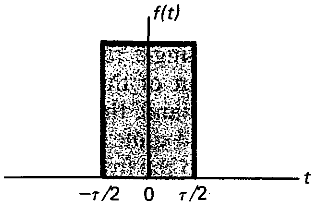


Fig. 36. A single rectangular pulse.

Hence

$$A(\omega) = 2U_0\tau \left|\frac{\sin(\omega\tau/2)}{\omega\tau/2}\right|, \qquad \varphi(\omega) = 0. \tag{8.24}$$

The amplitude spectrum $A(\omega)$ is shown in Fig. 37.

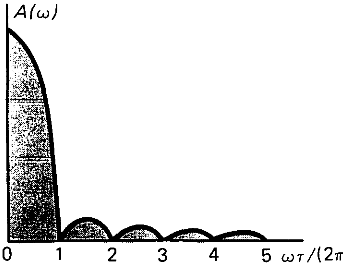


Fig. 37. Amplitude spectrum of a single rectangular pulse.

SPECTRUM OF AN EXPONENTIALLY DECREASING FUNCTION. Let us find the spectrum of the function

$$f(t) = \begin{cases} 0 & \text{for} \quad -\infty < t < 0, \\ U_0 e^{-t/\tau} & \text{for} \quad 0 < t < \infty. \end{cases}$$

The plot of this function is shown in Fig. 38. We have

$$F(\omega) = U_0 \int_0^\infty e^{-t/\tau - i\omega t}\,dt = U_0\tau/(1 + i\omega\tau).$$

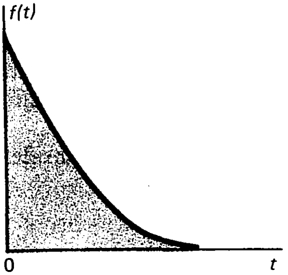


Fig. 38. Curve of an exponentially decreasing function.

Hence

$$A(\omega) = 2U_0\tau/\sqrt{1 + \omega^2\tau^2}, \qquad \varphi(\omega) = \omega\tau. \tag{8.25}$$

The amplitude spectrum $A(\omega)$ and the phase spectrum $\varphi(\omega)$ are shown in Figs. 39 and 40.

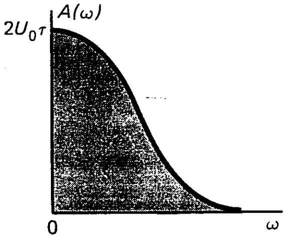


Fig. 39. Amplitude spectrum of an exponentially decreasing function.

RELATION BETWEEN THE PULSE DURATION AND THE SPECTRAL WIDTH. The pulse duration is the interval Δt of time during which the pulse differs significantly from zero. The spectral width is the interval $\Delta\upsilon$ of frequencies over which the spectral amplitude differs significantly from zero.

These definitions have an element of uncertainty, since it is not clear what we mean by the expression "differs significantly from zero". The relation between the duration of a particular pulse and the width of its spectrum undergoes a slight change depending on the definition of this concept. For a chosen definition of this concept, this ratio is different for pulses of different shapes. Hence there is no universal relation between the pulse duration and the spectral width. However, there is a universal regularity in the relations between the pulse duration and the spectral width, which is observed for different definitions of the concept "differs significantly from zero" and for different pulse shapes. According to this law, the spectral width is inversely proportional to the pulse duration. We shall derive this law by considering the example of a single rectangular pulse (see Fig. 36). In this case, the duration of the pulse is unequivocally known: $\Delta t = \tau$. On the other hand, the spectral width of this pulse (see Fig. 37) is also defined quite naturally: it is the interval of frequencies from zero up to the frequency at which the amplitude becomes equal to zero for the first time. This is so because the subsequent peaks of the amplitude are negligible in comparison to the principal peak at $\omega = 0$. Hence

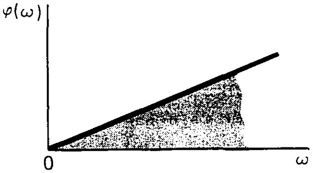


Fig. 40. Phase spectrum of an exponentially decreasing function.

we can write the following equality for the spectral width $\Delta\omega$:

$$\Delta\omega/2 = \pi. \tag{8.26}$$

Consequently, the spectral width over the frequencies $\Delta\upsilon = \Delta\omega/(2\pi)$ and the pulse duration $\Delta t = \tau$ are connected through the following relation:

$$\boxed{\Delta\upsilon\,\Delta t \simeq 1.} \tag{8.27}$$

The approximate equality instead of a strict equality has been used in this relation to emphasize the uncertainty in the definitions of $\Delta\upsilon$ and Δt. As a result, this relation is true only to within a multiplier of the order of unity. Relation (8.27) is used as a universal relation between the pulse duration and the spectral width.

The basic conclusion derived from (8.27) is that the shorter the duration of a pulse, the wider the spectrum of its frequencies. In other words, one cannot hope to represent a very short pulse through a set of harmonic functions with a small frequency interval. If $\Delta t \to 0$, the spectrum will contain all possible frequencies from very low to very high.

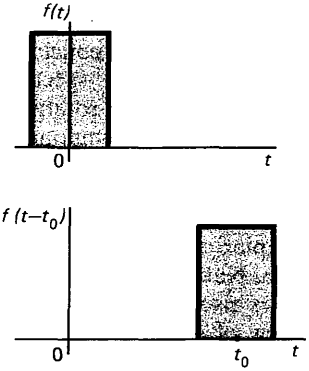


Fig. 41. Displacement of the time reference point.

DISPLACEMENT OF THE TIME REFERENCE POINT. Suppose that instead of the function $f(t)$ we have a function $f(t - t_0)$, where t_0 is a constant. This change in the argument does not change the shape of the pulse, but simply changes the reference point for measuring time (Fig. 41). This leads to a change in the Fourier transform of the function:

$$F_{t_0}(\omega) = \int_{-\infty}^{\infty} f(t - t_0)\,\mathrm{e}^{-i\omega t}\,\mathrm{d}t = \mathrm{e}^{-i\omega t_0} \int_{-\infty}^{\infty} f(\xi)\,\mathrm{e}^{-i\omega\xi}\,\mathrm{d}\xi$$

$$= F(\omega)\,\mathrm{e}^{-i\omega t_0}, \tag{8.28}$$

where we have changed over to a new integration variable $\xi = t - t_0$. Thus, a shift in the time reference point to the point t_0 changes the Fourier transform by $-\omega t_0$, i.e. leaves the amplitude spectrum unaltered but changes the phase spectrum.

FREQUENCY DISPLACEMENT OF SPECTRUM. Similarly, we can determine the influence of frequency shift, i.e. of the substitution $F(\omega) \to F(\omega - \omega_0)$:

$$f_{\omega_0}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega - \omega_0)\,\mathrm{e}^{i\omega t}\,\mathrm{d}\omega$$

$$= \frac{1}{2\pi}\mathrm{e}^{i\omega_0 t} \int_{-\infty}^{\infty} F(\xi)\,\mathrm{e}^{i\xi t}\,\mathrm{d}\xi = f(t)\,\mathrm{e}^{i\omega_0 t}, \tag{8.29}$$

where $\xi = \omega - \omega_0$. Thus, the displacement of the spectrum by ω_0 is equivalent to the modulation of the time function by a harmonic multiplier of frequency ω_0.

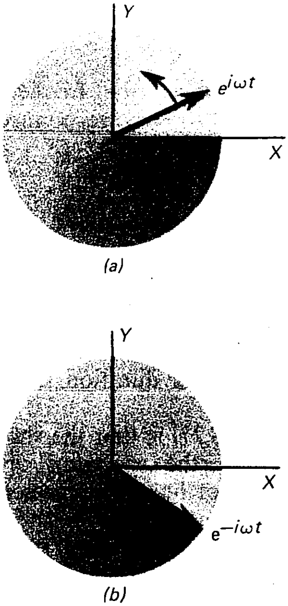


Fig. 42. Complex unit vectors with positive (*a*) and negative (*b*) directions of rotation.

NEGATIVE FREQUENCIES. The complex spectrum (8.8) completely defines the amplitude spectrum $A(\omega)$ as well as the phase spectrum $\varphi(\omega)$ through relations (8.16) and (8.17). In most cases, however, it is more convenient to analyze the spectrum of a function by directly using the expression for $F(\omega)$ without resorting to the quantities $A(\omega)$ and $\varphi(\omega)$. Since the argument of $F(\omega)$ assumes positive as well as negative values, this raises the question on the meaning of negative frequencies.

It should be noted that $e^{i\omega t}$ describes a complex unit vector (Fig. 42*a*) drawn from the origin of coordinates and rotating about this origin from the *X*-axis to the *Y*-axis with increasing *t*. This direction is taken as the positive direction and is connected with the *Z*-axis through the right-hand screw rule. With increasing *t*, the complex unit vector $e^{-i\omega t}$ rotates in the negative direction (Fig. 42*b*). Since the function $f(t)$ is expressed in terms of the Fourier transform by formula (8.7), we conclude that for $\omega > 0$, $F(\omega)$ describes the density of the spectral component of frequency ω with a positive direction of rotation, while $F(-\omega)$ describes the density of the spectral component with the same frequency but rotating in the opposite direction. Thus, a conversion to negative frequencies is associated with a change in the basis functions used in the Fourier transformation, i. e. with a change over to rotating complex vectors as the basis functions of the Fourier transformation. Naturally, all that has been said about the negative frequencies also remains valid for Fourier series in complex form in view of relation (8.7).

PARSEVAL'S THEOREM. Proceeding from representation (8.3) of a periodic function in the form of a Fourier series, let us determine the integral of $|f|^2$ over a period:

$$\int_{-T/2}^{T/2} f(t) f^*(t)\,\mathrm{d}t = \sum_{n,n'=-\infty}^{\infty} c_n c_{n'}^* \int_{-T/2}^{T/2} \exp[i(n-n')\omega t]\,\mathrm{d}t.$$

Since $\int_{-T/2}^{T/2} \exp[i(n-n')\omega t]\,\mathrm{d}t = T\delta_{nn'}$, we obtain the equality

$$\frac{1}{T}\int_{-T/2}^{T/2} |f(t)|^2\mathrm{d}t = \sum_{n=-\infty}^{\infty} |c_n|^2, \tag{8.30}$$

which is called Parseval's theorem. For the Fourier series (8.1) in real form, we obtain

$$\frac{1}{T}\int_{-T/2}^{T/2} |f|^2\mathrm{d}t = \frac{a_0^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2) = \frac{A_0^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty} A_n^2. \tag{8.31}$$

!

The coefficients of Fourier series or the Fourier transform of a function depend on the instant which serves as the reference point for measuring time.
The spectral composition of a function is uniquely determined by the Fourier series coefficients.
The amplitude and phase spectra are described by real numbers. The displacement of the time reference point is equivalent to the multiplication of the Fourier transform of a function by a phase multiplier which is a function of the displacement and frequency.

The frequency shift of a spectrum is equivalent to the multiplication of the function by a phase multiplier which is a function of the frequency shift and time.

PLANCHEREL'S THEOREM. If $f(t)$ is presented in the form (8.7), we obtain the following formula for the integral of the square of the modulus $|f(t)|^2$ in the same way as above:

$$\int_{-\infty}^{\infty} |f(t)|^2 \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 \mathrm{d}\omega = \int_{-\infty}^{\infty} |F(2\pi\nu)|^2 \mathrm{d}\nu . \tag{8.32}$$

This relation is called Plancherel's theorem.

SEC. 9. NATURAL LINEWIDTH OF RADIATION

The emission spectrum of a single stationary atom is calculated.

CLASSICAL MODEL OF A RADIATOR. The simplest classical model of a radiator is an electron performing harmonic oscillations about its equilibrium position. The equation for the free oscillations of the electron has the form

$$m\ddot{x} + m\omega_0 x = 0 , \tag{9.1}$$

where $m = 9.1 \times 10^{-31}$ kg is the electron mass, and ω_0 is the circular frequency of oscillations, its value for the optical spectral region being of the order of 10^{-15} s^{-1}. The solution of Eq. (9.1) can be written in the form

$$x(t) = A \cos(\omega_0 t - \varphi) , \tag{9.2}$$

where A is the amplitude of oscillations.

The total energy of an oscillating electron is given by

$$W = mx^2/2 + m\omega_0^2 x^2/2 = (m\omega_0^2/2) A^2 . \tag{9.3}$$

It is well known that the radiating power of an accelerated charge is

$$P = \frac{1}{6\pi\varepsilon_0} \frac{e^2}{c^3} \ddot{x}^2 . \tag{9.4}$$

Since the electron loses energy by radiation, its energy decreases and the oscillations are damped. Taking into account the radiation losses, we can write the law of conservation of energy for an oscillating electron in the form

$$\frac{\mathrm{d}W}{\mathrm{d}t} = -\langle P \rangle = -\frac{1}{12\pi\varepsilon_0} \frac{e^2}{c^3} \omega_0^4 A^2 , \tag{9.5}$$

where $\langle P \rangle$ is the value averaged over the period of oscillations, calculated from (9.2) and (9.4) by considering that the decrease in the amplitude over a period is small in comparison with the amplitude or, in other words, the energy emitted over a period is small in comparison with the total energy of radiation. This condition is always satisfied quite well.

Substituting the expression for A from (9.3) into (9.5), we obtain

$$\frac{dW}{dt} = -\gamma W, \tag{9.6}$$

where

$$\boxed{\gamma = \frac{1}{6\pi\varepsilon_0}\frac{e^2\omega_0^2}{mc^3}.} \tag{9.7}$$

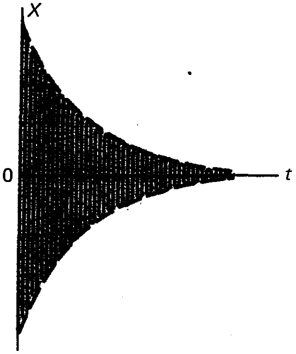


Fig. 43. Damping oscillations of an electron.

The solution of Eq. (9.6) has the form

$$W(t) = W_0 e^{-\gamma t}, \tag{9.8}$$

where $W_0 = W(0)$. Hence, in accordance with (9.3), the amplitude of electron oscillations varies according to the law

$$A = A_0 e^{-\gamma t/2}, \tag{9.9}$$

while the deviation x of the electron from its equilibrium position is given by the following equation instead of (9.2):

$$x(t) = A_0 e^{-\gamma t/2}\cos(\omega_0 t - \varphi). \tag{9.10}$$

It should be noted that all these calculations have been made under the assumption that the amplitude damping over an oscillation period $T = 2\pi/\omega_0$ is small:

$$\gamma T/2 = \pi\gamma/\omega_0 \ll 1. \tag{9.11}$$

Thus, the condition that the damping be small has the form

$$\gamma \ll \omega_0. \tag{9.12}$$

Substituting the values of the constants ε_0, e, m and c into (9.7), we find that condition (9.12) is always satisfied well for optical frequencies ($\omega_0 \sim 10^{15}$ s^{-1}). Under typical conditions, the amplitude is reduced to half over several million oscillations. In other words, $\gamma \sim 10^8$ s^{-1}. A graphical representation of such a slow change in amplitude is difficult, hence the electron oscillations are shown in Fig. 43 for a much higher (by a factor of several hundred thousands) value of γ.

Taking into account the damping of oscillations, we can replace (9.1) by the equation

$$m\ddot{x} + m\gamma\dot{x} + m\omega_0^2 x = 0. \tag{9.13}$$

A direct verification shows that for $\gamma \ll \omega_0$, the solution of (9.13) is identical to (9.10).

SPECTRAL COMPOSITION OF RADIATION. If an electron begins to oscillate and emit radiation at the instant

$t = 0$, the displacement $x(t)$ can be written in the form

$$x(t) = \begin{cases} 0 \text{ for } t < 0, \\ e^{-\gamma t/2}(A_0 e^{i\omega_0 t} + A_0^* e^{-i\omega_0 t}) \text{ for } t > 0, \end{cases} \tag{9.14}$$

since $x(t)$ is a real quantity.

The energy emitted in the interval $0 < t < \infty$ is given by

$$W = -\int_0^\infty \frac{dW}{dt} dt = \frac{1}{6\pi\varepsilon_0} \frac{e^2}{c^3} \int_{-\infty}^{\infty} \ddot{x}^2 dt, \tag{9.15}$$

where we have used formula (9.4) and considered that $x = 0$ for $t < 0$.

Let us write down the frequency distribution of the radiant energy (9.15). Since the damping is small ($\gamma \ll \omega_0$), it can be assumed that

$$\ddot{x} \approx -\omega_0^2 x, \tag{9.16}$$

while $x(t)$ can be presented in the form of a Fourier integral:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} d\omega \int_{-\infty}^{\infty} x(\xi) e^{-i\omega\xi} d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega, \tag{9.17}$$

where we have used the notation

$$F(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt = \int_0^\infty x(t) e^{-i\omega t} dt \tag{9.18}$$

considering that $\ddot{x} = 0$ for $t < 0$.

Substituting $\ddot{x}$ from (9.16) into (9.15) and expressing $x(t)$ through formula (9.17), we obtain

$$W = \frac{1}{6\pi\varepsilon_0} \frac{e^2}{c^3} \omega_0^4 \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \int_{-\infty}^{\infty} F(\xi) e^{i\xi t} d\xi = \frac{1}{24\pi^3\varepsilon_0} \frac{e^2}{c^3} \omega_0^4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\omega) F(\xi) d\omega \, d\xi \int_{-\infty}^{\infty} e^{i(\omega+\xi)t} dt. \tag{9.19}$$

Taking into account the relation

$$\int_{-\infty}^{\infty} e^{i(\omega+\xi)t} dt = 2\pi\delta(\omega + \xi), \tag{9.20}$$

we can present (9.19) in the form

$$W = \frac{1}{12\pi^2\varepsilon_0} \frac{e^2}{c^3} \omega_0^4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\omega) F(\xi) \delta(\omega + \xi) d\omega \, d\xi$$

$$= \frac{1}{12\pi^2\varepsilon_0}\frac{e^2}{c^3}\omega_0^4 \int_{-\infty}^{\infty} F(\omega)F(-\omega)\,d\omega$$

$$= \frac{1}{6\pi^2\varepsilon_0}\frac{e^2}{c^3}\omega_0^4 \int_{0}^{\infty} F(\omega)F(-\omega)\,d\omega$$

$$= \int_0^\infty w(\omega)\,d\omega, \tag{9.21}$$

where we have considered that $F(\omega)F(-\omega)$ is an even function of ω, and introduced the notation

$$w(\omega) = \frac{1}{6\pi^2\varepsilon_0}\frac{e^2}{c^3}\omega_0^4 F(\omega)F(-\omega). \tag{9.22}$$

The function $w(\omega)$ which is the density of the wave energy corresponding to the frequency ω, characterises the frequency distribution of energy. Substituting (9.14) into (9.18), we obtain

$$F(\omega) = \int_0^\infty A_0\exp\{[-\gamma/2 + i(\omega_0 - \omega)]\,t\}\,dt$$

$$+ \int_0^\infty A_0^*\exp\{[-\gamma/2 - i(\omega_0 + \omega)]\,t\}\,dt$$

$$= \frac{A_0}{\gamma/2 - i(\omega_0 - \omega)} + \frac{A_0^*}{\gamma/2 + i(\omega_0 - \omega)}. \tag{9.23}$$

The general form of the expression for $F(\omega)F(-\omega)$ appearing in (9.22) turns out to be quite cumbersome. Under the condition $\gamma \ll \omega_0$, however, the terms on the right-hand side of (9.23) differ considerably in their value at the frequencies $\omega \approx \omega_0$. The term containing the difference $\omega_0 - \omega$ is much larger than the term containing the sum $\omega_0 + \omega$. Hence it is sufficient to retain just the first term containing A_0 for $F(\omega)$ and the second term with A_0^* for $F(-\omega)$. Thus, we can write

$$F(\omega)F(-\omega) = \frac{A_0}{\gamma/2 - i(\omega_0 - \omega)}\,\frac{A_0^*}{\gamma/2 + i(\omega_0 + \omega)}$$

$$= \frac{A_0A_0^*}{(\omega_0 - \omega)^2 + (\gamma/2)^2}. \tag{9.24}$$

Formula (9.22) then assumes the form

$$w(\omega) = \frac{1}{6\pi^2\varepsilon_0}\frac{e^2}{c^3}\omega_0^4\frac{A_0A_0^*}{(\omega_0 - \omega)^2 + (\gamma/2)^2}. \tag{9.25}$$

?

Does a change in the time reference point cause a shift in the amplitude (a) or phase (b) spectrum? What is the complex spectrum of a function? How is it related to the amplitude and phase spectra?

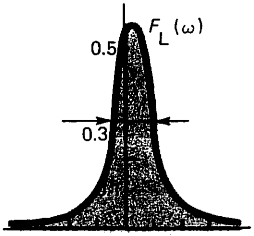


Fig. 44. Lorentz shape of a line.

The peak of the emission intensity lies near the frequency $\omega = \omega_0$. As we move away from this frequency, the intensity drops rapidly. The volume density of radiation at the frequencies

$$\omega_2 = \omega_0 + \gamma/2 \text{ and } \omega_1 = \omega_0 - \gamma/2, \tag{9.26}$$

is half its value at the frequency ω_0. This means that the main part of the radiant energy falls within the frequency interval

$$\delta\omega = \omega_2 - \omega_1 = \gamma, \tag{9.27}$$

called the width of the emission line.

LORENTZ SHAPE AND WIDTH OF EMISSION LINE. It is convenient to write formula (9.22) in the form

$$w = 2mA_0A_0^*\omega_0^2F_L(\omega), \tag{9.28}$$

where

$$F_L(\omega) = \frac{\gamma/(2\pi)}{(\omega_0 - \omega)^2 + (\gamma/2)^2} \tag{9.29}$$

is the normalized function for the Lorentz shape of the emission line (Fig. 44). The function (9.29) satisfies the normalization condition

$$\int_{-\infty}^{\infty} F_L(\omega)\,\mathrm{d}\omega = 1. \tag{9.30a}$$

The main contribution to the integral comes from the domain of integration near $\omega \approx \omega_0$. The contribution to the integral from the domain with $\omega < 0$ has the order of magnitude of the terms which were neglected while calculating $F(\omega)\,F(-\omega)$ from formula (9.24) derived from (9.23). Hence, to the accuracy with which $F(\omega)\,F(-\omega)$ is expressed in (9.24), we can write normalization condition (9.30a) in the form

$$\int_{0}^{\infty} F_L(\omega)\,\mathrm{d}\omega = 1. \tag{9.30b}$$

The Lorentz shape of an emission line is obtained under "natural" conditions of emission, when radiation damping is the only factor influencing the radiation emitted by an oscillator. Hence this shape is often called the natural shape of the emission line, and the line width is called the γ-natural line width.

The natural width of the emission line is very small. For optical frequencies, $\gamma/\omega_0 \sim 10^{-7}$. If in the analysis of a certain problem the variation of frequency by about 1/10 000 000 of the frequency ω_0 is insignificant, it can be assumed that the radiation is emitted entirely at the frequency ω_0 and the radiation can be considered as monochromatic. Such an approximation is valid in most cases considered in optics. Only in some cases we have to take into account the finiteness of the natural width of emission.

EMISSION TIME. It can be seen from (9.14) that the oscillation amplitude of an electron decreases by a factor e = 2.7 over a time

$$\boxed{\tau = 2/\gamma = 2/(\delta\omega),} \tag{9.31}$$

called the emission time. It is clear that the smaller the emission time, the larger the emission linewidth. A strictly monochromatic radiation is possible only for an infinitely long emission time.

SHAPE OF ABSORPTION LINE. The equation of motion for an elastically bound electron in an electric field E has the form

$$m\ddot{x} + m\gamma\dot{x} + m\omega_0^2 x = eE, \tag{9.32}$$

where it is assumed that the field strength E is directed along the X-axis. If

$$E(t) = E_0 e^{i\omega t}, \tag{9.33}$$

then in the steady state $x(t)$ is a harmonic function with the same frequency ω:

$$x(t) = x_0 e^{i\omega t}. \tag{9.34}$$

Since $\dot{x} = i\omega x$, $\ddot{x} = -\omega^2 x$, Eq. (9.32) assumes the form

$$m(-\omega^2 + i\omega\gamma + \omega_0^2)x = eE, \tag{9.35}$$

whence

$$x = \frac{e}{m}\frac{E}{\omega_0^2 - \omega^2 + i\omega\gamma}. \tag{9.36}$$

An oscillating electron is a source of radiation, and the energy is transmitted to it by the incident wave. As a result, the energy is absorbed by the elastically bound electron. The absorption intensity is proportional to the square of the amplitude (9.36) of the oscillating electron, i.e. xx^*. Hence if a radiation with a continuous frequency spectrum is incident on an elastically bound electron, an absorption line is formed, its shape being given by the expression

$$\Phi(\omega) = \text{const}\, xx^* = \frac{\text{const}}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2}. \tag{9.37}$$

For $\omega_0 \ll \gamma$, this function is nonzero only near the frequency ω_0, when the denominator in (9.37) can be transformed as follows:

$$(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2 = (\omega_0 - \omega)^2(\omega_0 + \omega)^2$$
$$+ \omega^2\gamma^2 \approx 4\omega_0^2(\omega_0 - \omega)^2 + \omega_0^2\gamma^2$$
$$= 4\omega_0^2[(\omega_0 - \omega)^2 + (\gamma/2)^2]. \tag{9.38}$$

Relation (9.37) then assumes the form

$$\Phi(\omega) = \frac{\text{const}}{(\omega_0 - \omega)^2 + (\gamma/2)^2}. \tag{9.39}$$

Hence the absorption line has Lorentz shape with the same width as the emission line.

QUANTUM INTERPRETATION OF THE SHAPE OF EMISSION LINE. In quantum theory, radiation is interpreted as the emission of photons. The frequency ω of a photon is connected with its energy E through the relation

$$E = \hbar\omega. \tag{9.40}$$

The emission of the photon takes place as a result of a transition of the emitting atom from the state with energy E_n to the state with energy E_m, so that $E = E_n - E_m$. Consequently, the emission frequency

$$\omega_{nm} = (E_n - E_m)/\hbar \tag{9.41}$$

has quite a definite value. However, photons of different frequencies are emitted as a result of a quantum transition, since the energy E_n of the excited state does not have a definite value. In other words, the excited energy level has a finite width. In quantum theory, the emission time is interpreted as the time during which the atom is in an excited state, while formula (9.31) simply states that the width of an energy level is inversely proportional to the lifetime of an electron on this level. The electron can remain for an infinitely long time in the ground state with the lowest energy if the atom is isolated. Hence the width of the ground energy level is equal to zero. If the emission line is formed due to the transitions of the electron from a certain excited state to the ground state, the Lorentz shape of the line reflects the energy distribution of electrons on the excited energy level. If the atom is in a radiation field or interacts with other atoms, it can pass from the ground state to an excited state. Consequently, it has a finite lifetime in the ground state, and hence the ground state is not infinitely narrow, but rather has a finite width. This leads to corresponding broadening of the emission line.

!

Natural linewidth is attained only in the radiation of a stationary atom isolated from the surroundings. According to classical interpretation, the natural width of an emission line is due to the finite duration of the emission process. According to the quantum theory, the natural width of an emission line is due to the finite width of energy levels.

?

How are the shapes of emission and absorption lines related? How is the shape of an absorption line associated with the resonance amplitude characteristic of a damped linear oscillator?
Why can we not assign the frequency spectrum corresponding to the natural shape of an emission line to an individual photon?

QUASI-MONOCHROMATIC WAVE. If the width of the emission line is quite small, the wave can be represented in form (2.39), where ω is the frequency at the centre of the line. In other words, we can assume that the wave is monochromatic with a frequency equal to the average frequency of the nonmonochromatic wave. This is possible under the condition

$$\Delta\omega \ll \omega, \tag{9.42}$$

called the quasi-monochromaticity condition.

A quasi-monochromatic wave can be treated as a monochromatic wave over a period smaller than the coherence time (see Secs. 13, 26, 30). If in some phenomenon the wave variation has to be considered over a period longer than the coherence time, it is not possible to prevent it in the form of a monochromatic wave of type (2.39).

SEC. 10. SPECTRAL LINE BROADENING

The broadening of spectral lines is classified. The Doppler and collision broadenings are described.

REASONS FOR BROADENING. The natural shape of an emission line appears under ideal conditions when the emitting atom is at rest and is not subjected to the action of any external forces during the process of emission. In practice, ideal conditions are never realized completely because of the thermal motion of the atom and its interaction with other atoms. These circumstances generally affect the shape of the emission line. Hence the natural shape of a line is usually not observed under actual conditions.

UNIFORM AND NONUNIFORM BROADENING. The factors leading to the broadening of emission lines can be divided into two groups. One group causes the same variation in the emission line of each atom. Such a broadening of lines is called uniform.

The other group of factors causes different types of variation in the emission lines of individual atoms. In this case, the line broadening observed in the emission of an aggregate of atoms is the net result of different variations in the emission lines of individual atoms. Such a broadening of lines is called nonuniform broadening.

NATURAL WIDTH OF EMISSION LINE AS A UNIFORM BROADENING. The natural width of emission lines is an example of the uniform broadening of lines since this broadening is the same in the emission of all atoms of a given species and is determined only by the emission time. The observed shape of the emission line for a set of atoms coincides with the shape of the emission line for an individual atom.

COLLISION BROADENING. In a gas under normal pressure and room temperature, the time τ_0 over which the emission from an atom is not violated by interactions is of the order of 10^{-11} s. Considering that the time τ of natural emission in (9.31) is of the order of 10^{-8} s, the atom experiences over one hundred violations of the emission process in the course of emission. Each such violation causes a change in the emission conditions and is equivalent to an interruption. After this, the emission again continues with the same frequency ω_0. This means that the phase of the emitted wave undergoes a random variation at the instant of violation of the emission conditions. The entire process of emission is as if split into individual acts of emission whose duration is equal to the interval of time between two successive violations of the emission conditions. This leads to a broadening of the emission line. Since the time between two violations is about one-hundredth of the total emission time, the width of the lines under these conditions is found to be about 100 times the natural width of the emission line. This broadening is called collision broadening since it is a consequence of the collisions of atoms.

To calculate the collision broadening, let us assume that the natural width of the emission line is equal to zero. Since the collisions are random, the time between successive collisions obeys Poisson's distribution, i.e. the probability that the time between two successive collisions lies in the interval between ξ and $\xi + d\xi$ is given by

$$\boxed{p(\xi)\,d\xi = (1/\tau_0)\exp(-\xi/\tau_0)\,d\xi,} \tag{10.1}$$

where τ_0 is the mean time between collisions. The damping of the amplitude of electron oscillations, and hence of the amplitudes of the field of emitted waves, is insignificant over the period between two collisions. Hence in the process of emission between collisions at the instants t_0 and $t_0 + \xi$, the amplitude of the electric field strength of the emitted wave is practically constant:

$$E(t) = E_0 \exp(i\omega_0 t + i\varphi)\,(t_0 < t < t_0 + \xi), \tag{10.2}$$

where φ is the random phase.

The spectral composition of the electric field of the wave in the interval of time under consideration is given by the formula

$$E_\omega = \int_{-\infty}^{\infty} E(t)\,e^{-i\omega t}dt = \int_{t_0}^{t_0+\xi} E_0 \exp(i\omega_0 t - i\omega t + i\varphi)\,dt$$

$$= E_0 \exp[i(\omega_0 - \omega)t_0 + i\varphi]\,\frac{\exp[i(\omega_0 - \omega)\xi] - 1}{i(\omega_0 - \omega)}. \tag{10.3}$$

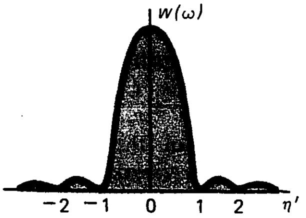


Fig. 45. Spectral composition of the emissive power between successive collisions [$\eta' = (\omega - \omega_0)\tau/(2r)$].

Hence the spectral composition of the emissive power $w(\omega)$ averaged over the oscillation period (Fig. 45) is given by the relation

$$w_1(\omega) \propto E_\omega E_\omega^* = 4E_0^2 \frac{\sin^2[(\omega_0 - \omega)\xi/2]}{(\omega_0 - \omega)^2}. \qquad (10.4)$$

Formula (10.4) describes the intensity of radiation from an individual atom over an interval of time ξ. Other atoms also emit radiation at the same time and their mean free times are distributed in accordance with (10.1). Hence to find the spectral composition of the total intensity of radiation from all atoms, we must average (10.4) over ξ, taking (10.1) into account:

$$w(\omega) \sim \frac{1}{\tau_0}\int_0^\infty \frac{\sin^2[(\omega_0 - \omega)\xi/2]}{(\omega_0 - \omega)^2} e^{-\xi/\tau_0}\, d\xi$$
$$= \frac{1}{2}\frac{1}{(\omega_0 - \omega)^2 + (1/\tau_0)^2}. \qquad (10.5)$$

This means that the collision broadening leads to the Lorentz shape of the line with a width $\gamma_{\mathrm{br}} = 2/\tau_0$.

Collision broadening is uniform since it is characterized by the average interval of time τ between collisions which are identical for all atoms and there is no experimental means of attributing the emission of a certain frequency to a particular group of atoms.

Although the collision width is two orders of magnitude more than the natural width, its absolute value is negligibly small in most cases and we can assume that we are considering radiation of a single frequency ω_0.

DOPPLER BROADENING. If an atom is moving during the emission process, the frequency of the radiation emitted by it is not equal to the frequency of radiation emitted by an atom at rest. The laws of conservation of energy and momentum for a radiating atom have the form

$$E_2 + (1/2)Mv_2^2 = E_1 + (1/2)Mv_1^2 + \hbar\omega, \qquad (10.6)$$

$$M\mathrm{v}_2 = M\mathrm{v}_1 + \hbar\mathrm{k}, \qquad (10.7)$$

where E_2 is the energy of the excited state of the atom, E_1 is the energy of the lower level to which the emitting electron undergoes a transition, M is the mass of the atom, v_2 and v_1 represent the electron velocity before and after the act of emission, while $\hbar\omega$ and $\hbar\mathrm{k}$ are the energy and momentum of the emitted photon. Denoting by ω_0 the frequency of the radiation emitted by an atom which remains at rest before and after the act of emission, we can write

$$\hbar\omega_0 = E_2 - E_1. \tag{10.8}$$

Squaring (10.7) and eliminating E_1, E_2 and v_1 from (10.6)-(10.8), we arrive at the equation

$$\hbar(\omega_0 - \omega) = \hbar^2 k^2/(2M) - \hbar \mathrm{v}_2 \cdot \mathrm{k}. \tag{10.9}$$

Directing the Z-axis of the Cartesian system of coordinates along the wave vector k of the emitted photon and considering that $k = \omega/c$, we can write (10.9) in the form

$$\omega_0 - \omega = \hbar\omega^2/(2Mc^2) - \omega v_{2z}/c. \tag{10.10}$$

The atoms have nonrelativistic velocities and $v_{2z}/c \sim 10^{-5}$ at room temperature, while for $\omega \sim 10^{15}\,\mathrm{s}^{-1}$ and typical atomic masses, the value of $k\omega^2/(2Mc^2)$ is of the order of 10^{-9}. Hence the minuend on the right-hand side of (10.10) can be neglected in comparison with the subtrahend. This gives

$$\omega = \frac{\omega_0}{1 - v_{2z}/c} \approx \omega_0\left(1 + \frac{v_{2z}}{c}\right). \tag{10.11}$$

This formula describes the Doppler effect: in the direction of the velocity of the atom, the frequency of the emitted photon increases by $\Delta\omega = \omega v/c$, while the frequency of a photon emitted in the opposite direction decreases by the same amount.

According to Maxwell's distribution, the fraction of atoms whose velocities lie between v_{2z} and $v_{2z} + \mathrm{d}v_{2z}$ is proportional to

$$\mathrm{d}v_{2z} \exp[-Mv_2^2/(2kT)], \tag{10.12}$$

where k is the Boltzmann constant. From (10.11), we obtain

$$v_{2z} = (\omega - \omega_0)\,c/\omega_0, \quad \mathrm{d}v_{2z} = (c/\omega_0)\,\mathrm{d}\omega. \tag{10.13}$$

Then (10.12) leads to the following frequency distribution of the radiant energy

$$\underline{\overline{w(\omega)\,\mathrm{d}\omega \sim \exp[-Mc^2(\omega - \omega_0)^2/(2\omega_0^2 kT)]\,\mathrm{d}\omega.}} \tag{10.14}$$

The shape of the emission line described by this formula is called Gaussian. The peak of intensity for a Gaussian emission line lies at the frequency $\omega = \omega_0$. The width Δ of a Gaussian line is the separation between the frequencies corresponding to half the peak intensity. These frequencies ω_1 and ω_2 are obtained as the roots of the equation

$$\underline{\overline{1/2 = \exp[-Mc^2(\omega - \omega_0)^2/(2\omega_0^2 kT)].}} \tag{10.15}$$

!

The uniform broadening of a line observed in the emission of an aggregate of atoms is due to the identical broadening of emission lines of individual atoms. The shape of a uniformly broadened line is the same as the shape of the broadened emission lines of individual atoms. The nonuniform broadening of lines observed in the emission of an aggregate of atoms is due to different variations in the emission lines of individual atoms.

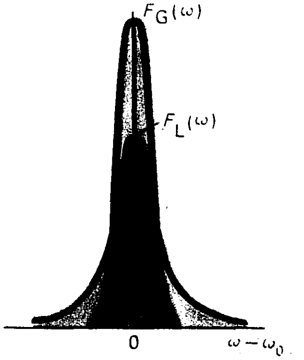


Fig. 46. Gaussian and Lorentz emission lines having the same width and normalized to unity.

Hence

$$\Delta = 2\omega_0\,[2kT\ln 2/(Mc^2)]^{1/2}. \tag{10.16}$$

It is convenient to present Gaussian distribution in the form

$$F_G(\omega) = [1/(\sigma\sqrt{2\pi})]\exp[-(\omega-\omega_0)^2/(2\sigma^2)], \tag{10.17}$$

where

$$\sigma = [kT\omega_0^2/(Mc^2)]^{1/2} = \Delta/(2\sqrt{2\ln 2}). \tag{10.18}$$

In such a form, the Gaussian distribution is normalized to unity:

$$\int_{-\infty}^{\infty} F_G(\omega)\,d\omega = 1. \tag{10.19}$$

Figure 46 shows a Lorentz line (9.29) and a Gaussian line (10.17) having the same width $\gamma = \Delta$. The area under these curves is equal to unity.

Under normal pressure and at room temperature, the Doppler width of a line is more than the natural width by about two orders of magnitude. Under such conditions, it has the same order as the width of a line formed by collision broadening.

SHAPE OF A COMPOSITE EMISSION LINE. When an emission line undergoes broadening due to the simultaneous action of several factors, the resultant line is connected with the components through a simple formula under the assumption that all the lines have the same central frequency ω_0. If $F_1(\omega)$ and $F_2(\omega)$ characterize two emission lines with the same central frequency ω_0, the shape of the composite emission line is defined as follows:

$$F(\omega) = \int_{-\infty}^{\infty} F_1(\xi)F_2(\omega+\omega_0-\xi)\,d\xi. \tag{10.20}$$

The central frequency of the composite line is the same as that of the components.

The Lorentz and Gaussian lines have an important peculiarity. The addition of two Lorentz lines with width γ_1 and γ_2 leads to a Lorentz line with width $\gamma = \gamma_1 + \gamma_2$. The addition of two Gaussian lines of widths Δ_1 and Δ_2 results in a Gaussian line of width $\Delta = \sqrt{\Delta_1^2 + \Delta_2^2}$.

!

The shape of a nonuniformly broadened emission line for an aggregate of atoms does not coincide with the shape of nonuniformly broadened emission lines for individual atoms. The shape of a collision-broadened line is Lorentzian, while that of a Doppler-broadened line is Gaussian.

?

Why is collision broadening uniform?
Why is Doppler broadening not uniform?
What is the order of magnitude of Doppler and collision broadenings of lines in a gas under normal conditions?
Is the frequency of atomic collisions leading to the collision broadening of lines equal to the gasdynamic frequency of collisions? Why? What is the meaning of collision frequency in these two cases?

SEC. 11. MODULATED WAVES

The amplitude-, frequency- and phase modulation of waves is described.

MODULATION. The harmonic oscillation describing a wave is characterized by its amplitude, frequency and phase. The variation of these parameters in the process of oscillation is called modulation. The waves obtained as a result of modulation are called modulated waves.

AMPLITUDE MODULATION. Oscillations with modulated amplitude can be presented in the form

$$f(t) = [A_0 + a(t)] \cos \omega t, \tag{11.1}$$

where $a(t)$ describes the modulation, ω is the harmonic oscillation frequency, and $|a(t)| < A_0$.

To begin with, let us consider the case when $a(t)$ is a harmonic function

$$a(t) = a_0 \cos \Omega t, \tag{11.2}$$

with a frequency $\Omega \ll \omega$. In this case, Eq. (11.1) assumes the form

$$f(t) = (A_0 + a_0 \cos \Omega t) \cos \omega t. \tag{11.3}$$

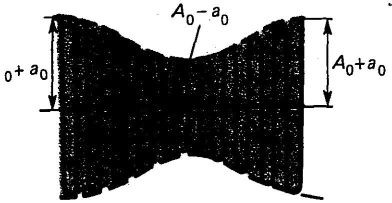


Fig. 47. Harmonic amplitude-modulated oscillation.

The plot of this oscillation is shown in Fig. 47. With the help of the formula for cosines of the sum and the difference of angles, we can write (11.3) in the form

$$f(t) = A_0 \cos \omega t + (1/2)\, a_0 \cos (\omega - \Omega)\, t + (1/2)\, a_0 \cos (\omega + \Omega)\, t, \tag{11.4}$$

whence it can be concluded that the spectral composition of the oscillation is reduced to three frequencies: ω, $\omega - \Omega$ and $\omega + \Omega$ (Fig. 48). The frequency ω is called the carrier frequency, while the frequencies $\omega \pm \Omega$ are called sideband frequencies.

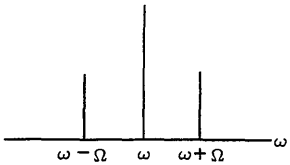


Fig. 48. Spectrum of harmonic amplitude-modulated oscillations.

If $a(t)$ is not a harmonic but a periodic function having a period $T = 2\pi/\Omega$, it can be presented in the form of a Fourier series in frequencies that are multiples of Ω. Substituting the Fourier series into formula (11.1) and transforming each term of the series after multiplying it by $\cos \omega t$ in the same way as was done for going over from (11.3) to (11.4), we obtain a series containing the frequencies ω and $\omega \pm n\Omega$ ($n = 1, 2, \ldots$). In other words, the spectrum consists of a set of frequencies that differ from one another by Ω on either side of the carrier frequency ω_0 (Fig. 49). The width of the spectrum is determined by the spectral width of the function $a(t)$. If $a(t)$ is a nonperiodic function which can be represented by a Fourier integral, its spectrum will be continuous. In this case, substituting the Fourier integral form of the function $a(t)$ into

!

Frequency and phase modulations are equivalent only if they are harmonic. In the case of anharmonic modulation, the modulation signals are quite different.

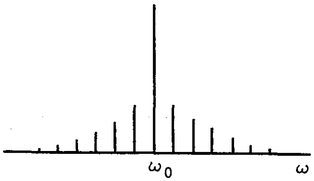


Fig. 49. Spectrum of periodic amplitude-modulated oscillations.

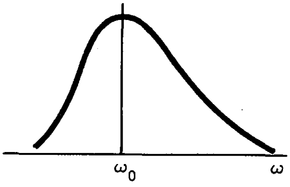


Fig. 50. Continuous spectrum of nonperiodic amplitude-modulated oscillations.

(11.1) and transforming the harmonic components in the same way as before, we obtain for the modulated oscillation a continuous spectrum extending on both sides of the carrier frequency ω_0 (Fig. 50). Thus, in this case also, the width of the spectrum is determined by the width of the spectrum of $a(t)$.

All that has been stated above is valid for oscillations, say, of the electric field strength at a fixed point in space. If the velocity of wave propagation is independent of frequency (nondispersive medium), the oscillation moves to another point in space (without any change in its shape) at the velocity of wave propagation. Hence the change in the shape of oscillation at a given point with time can be easily converted into the spatial shape of oscillations (see Sec. 5).

FREQUENCY- AND PHASE MODULATION. It is expedient to consider these two types of modulations together since the formulas describing them are closely related. The frequency-modulated and phase-modulated signals, however, differ considerably in structure.

The relation between frequency modulation and phase modulation is obtained by describing the oscillation phase in terms of the time-dependent (modulated) frequency according to the formula

$$\sin \Phi(t) = \sin \left[\int_0^t \omega(t)\, \mathrm{d}t\right], \tag{11.5}$$

where $\Phi(t)$ is the phase of the oscillation and $\omega(t)$ is the modulated frequency. The initial phase of oscillations is assumed to be equal to zero. It follows from the expression

$$\Phi(t) = \int_0^t \omega(t)\, \mathrm{d}t \tag{11.6}$$

that the instantaneous value of the frequency $\omega(t)$ is connected with the phase $\Phi(t)$ through the relation

$$\omega(t) = \mathrm{d}\Phi(t)/\mathrm{d}t. \tag{11.7}$$

With the help of (11.6) and (11.7), we can go over from formulas describing frequency modulation to formulas describing phase modulation.

Let us consider harmonic frequency modulation. In this case, we have

$$\omega(t) = \omega_0 + \Delta\omega \cos \Omega t, \tag{11.8}$$

where ω_0 is the constant frequency about which frequency oscillations with amplitude $\Delta\omega$ and frequency Ω take place. In accordance with (11.6) we have

$$\Phi(t) = \int_0^t \omega(t)\, \mathrm{d}t = \omega_0 t + (\Delta\omega/\Omega) \sin \Omega t. \tag{11.9}$$

?

What is the difference between the methods of realization of frequency and phase modulation?
How does the capacity of a transmission line depend on the carrier frequency?

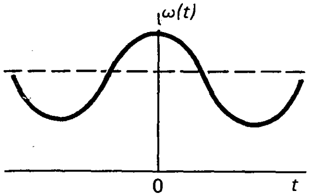


Fig. 51. Harmonic frequency modulation.

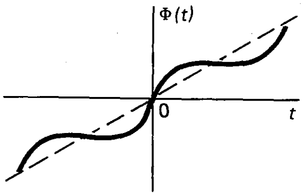


Fig. 52. Harmonic phase modulation.

This means that the phase $\Phi(t)$ is also modulated according to a harmonic law with the same frequency Ω and amplitude $\Delta\omega/\Omega$ of modulation. If the phase is modulated according to the harmonic law

$$\Phi(t) = \omega_0 t + \Delta\Phi \sin \Omega t, \tag{11.10}$$

then differentiating (11.10) with respect to time and taking into account (11.7) we arrive at the following relation

$$\omega(t) = \mathrm{d}\Phi/\mathrm{d}t = \omega_0 + \Delta\Phi\Omega \cos \Omega t. \tag{11.11}$$

This formula shows that the frequency is found to be modulated harmonically with the same modulation frequency Ω and amplitude $\Delta\Phi\Omega$. Figures 51 and 52 show the harmonic frequency modulation $\omega(t)$ and harmonic phase modulation $\Phi(t)$.

Thus, the frequency and phase modulations are exactly equivalent only when they are harmonic. Such an equivalence is not possible if the modulation is not harmonic, since in this case the frequency-modulated signals are quite different in shape from the phase-modulated signals. In the case of frequency modulation, large-scale phase oscillations correspond to slow signal variations, i. e. to low frequencies in the signal [$\Delta\Phi_{\max} = \Delta\omega/\Omega$ in (11.9)], while small phase oscillations correspond to rapid variations. On the contrary, small amplitudes of frequency oscillations correspond to slow variations of signal in the case of phase modulation ($\Delta\omega = \Delta\Phi\Omega$), while large oscillations correspond to rapid variations. The frequency- and phase modulations also differ in the manner of their realization. In the case of frequency modulation, the oscillation frequency of the generator is directly affected, while for the phase modulation the oscillation frequency of the generator is constant while the phase is modulated as the signal emitted by the generator propagates.

OSCILLATION SPECTRUM WITH HARMONIC FREQUENCY MODULATION. Let us carry out the spectral analysis of a frequency-modulated signal with a harmonic modulation

$$E(t) = E_0 \sin[\omega_0 t + (\Delta\omega/\Omega) \sin \Omega t]. \tag{11.12}$$

Assuming that the modulation amplitude is small ($\Delta\omega/\Omega \ll 1$), we can expand (11.2) into a Taylor series in $(\Delta\omega/\Omega)\sin\Omega t$ and confine ourselves to the first-order terms:

$$\begin{aligned} E(t) &\approx E_0 [\sin \omega_0 t + (\Delta\omega/\Omega) \sin \Omega t \cos \omega_0 t] \\ &= E_0 \{\sin \omega_0 t + [\Delta\omega/(2\Omega)] \sin(\omega_0 + \Omega) t \\ &\quad - [\Delta\omega/(2\Omega)]\} \sin(\omega_0 - \Omega) t. \end{aligned} \tag{11.13}$$

Thus, in the first approximation, the spectrum contains just the frequencies ω_0 and $\omega_0 + \Omega$, $\omega_0 - \Omega$. In other words, the spectrum is analogous to the spectrum of an amplitude-modulated signal with the same frequency. However, such a correspondence is valid only if the modulation depth is small. As $\Delta\omega/\Omega$ increases, other spectral components in the frequency-modulated signal also begin to play a significant role. Hence, in general, a signal whose frequency is harmonically modulated contains an infinite number of frequencies in its spectrum and thus differs basically from a signal undergoing a harmonic amplitude modulation. The frequency modulation also differs from amplitude modulation in that the spectrum of an amplitude-modulated oscillation is linearly connected with the spectrum of the signal, while this connection is nonlinear in the case of frequency modulation. The addition of a new frequency to the spectrum leads to the introduction of a corresponding frequency in the amplitude-modulated spectrum without any change in the amplitudes of the remaining frequencies. For frequency modulation, the addition of a new frequency leads not only to the appearance of many new frequencies in the modulated spectrum, but also to a change in the amplitudes of the existing oscillations.

For arbitrary values of $\Delta\omega/\Omega$, the oscillation spectrum (11.12) is expressed in terms of the Bessel function $J_n(x)$ with an integral index n and will not be considered here.

SEC. 12. WAVE PACKETS

The properties of wave packets and a quasi-plane wave are described.

WAVE PACKET FORMED BY TWO WAVES. Only in vacuum is the velocity of electromagnetic waves independent of the frequency and equal to the velocity of light. In a medium, the velocity of an electromagnetic wave is less than the velocity of light and depends on the frequency. The dependence of the wave velocity on frequency is called dispersion.

Let us consider the superposition of two waves having frequencies ω_1 and ω_2, and the corresponding wave numbers k_1 and k_2:

$$E_1(z,t) = E_0 \cos(\omega_1 t - k_1 z), \quad E_2(z,t) = E_0 \cos(\omega_2 t - k_2 z). \tag{12.1}$$

We assume that the two waves have identical polarization and propagate in the same direction. The phase velocity of a wave is determined from the condition

$$\omega t - kz = \text{const}. \tag{12.2}$$

Differentiating this equation with respect to time, we obtain

$$\underline{v_{\mathrm{ph}} = v = \mathrm{d}z/\mathrm{d}t = \omega/k} \tag{12.3}$$

(for the sake of simplicity, we shall henceforth omit the subscript "ph" indicating the phase velocity). We denote the phase velocity in vacuum by c. In general, the phase velocities of waves in (12.1) may be different. The electric field strength of the wave formed as a result of superposition is given by the formula

$$E(z,t) = E_1 + E_2 = 2E_0 \cos[(\omega_1 - \omega_2)t/2 - (k_1 - k_2)z/2] \cos[(\omega_1 + \omega_2)t/2 - (k_1 + k_2)z/2]. \tag{12.4}$$

The shape of such a wave is shown in Fig. 10. If there is no dispersion, the electric field strength of the wave has the form (4.7). The wave propagates in the direction of positive values of the Z-axis with the velocity of light without any change in its shape, while the envelope of the amplitudes shown by a dashed line in Fig. 10 also moves with the velocity of light.

GROUP VELOCITY. The superposition of two or more waves having different frequencies results in the formation of a group of waves, or a wave packet. The velocity of a group of waves, or the group velocity, is the velocity of the peak of the amplitude envelope of a group of waves. Differentiating condition (12.4) for the constancy of the phase of the wave amplitude envelope written in the form

$$(1/2)(\omega_1 - \omega_2)t - (1/2)(k_1 - k_2)z = \mathrm{const}, \tag{12.5}$$

with respect to t, we obtain the group velocity:

$$v_{\mathrm{g}} = \mathrm{d}z/\mathrm{d}t = (\omega_1 - \omega_2)/(k_1 - k_2). \tag{12.6}$$

If there is no dispersion, then $\omega_1 = ck_1$ and $\omega_2 = ck_2$. Hence we obtain from (12.6) $v_{\mathrm{g}} = c$, i.e. the group velocity is the same as the phase velocity. If there is dispersion, the group velocity differs from the phase velocity. As a result, the amplitude envelope and the component waves move with different velocities, thus causing a change in the shape of the envelope during wave propagation. In other words, in the presence of dispersion a wave packet propagation is accompanied by a change in its shape.

If the frequencies of the component waves are close in value ($\omega_2 \to \omega_1$), we obtain from (12.6) the following formula for group velocity:

$$\underline{v_{\mathrm{g}} = \mathrm{d}\omega/\mathrm{d}k.} \tag{12.7}$$

This formula is valid not only for two waves with infinitely close frequencies, but also for any wave packet formed by superposition of an infinite number of waves with close frequencies, since it is a differential formula.

• SUPERPOSITION OF OSCILLATIONS WITH EQUIDISTANT FREQUENCIES. Let there be N oscillations of the same amplitude E_0, their frequencies differing by $\delta\omega$. The result of superposition of these oscillations is given by the formula

$$E(t) = E_0 \cos \omega t + E_0 \cos(\omega + \delta\omega) t + E_0 \cos(\omega + 2\delta\omega) t + \ldots + E_0 \cos[\omega + (N-1)\delta\omega] t. \tag{12.8}$$

The summation of this series can best be carried out through an exponential representation of harmonic functions:

$$\begin{aligned} E(t) &= \mathrm{Re}\left[E_0 \sum_{n=0}^{N-1} \mathrm{e}^{i(\omega t + n\delta\omega t)} \right] \\ &= E_0 \mathrm{Re}\left[\mathrm{e}^{i\omega t} \sum_{n=0}^{N-1} \mathrm{e}^{in\delta\omega t} \right] \\ &= E_0 \mathrm{Re}\left[\mathrm{e}^{i\omega t} \frac{1 - e^{iN\delta\omega t}}{1 - \mathrm{e}^{i\delta\omega t}} \right] \\ &= E_0 \mathrm{Re}\left[\mathrm{e}^{i\omega t} \frac{\mathrm{e}^{iN\delta\omega T/2} (\mathrm{e}^{-iN\delta\omega t/2} - \mathrm{e}^{iN\delta\omega t/2})}{\mathrm{e}^{i\delta\omega t/2} (\mathrm{e}^{-i\delta\omega t/2} - \mathrm{e}^{i\delta\omega t/2})} \right] \\ &= E_0 \mathrm{Re}\left[\mathrm{e}^{i[\omega t + (N-1)\delta\omega t/2]} \frac{\sin(N\delta\omega t/2)}{\sin(\delta\omega t/2)} \right] \\ &= E_0 \frac{\sin(N\delta\omega t/2)}{\sin(\delta\omega t/2)} \cos\langle \omega \rangle t, \end{aligned} \tag{12.9}$$

where $\langle \omega \rangle = \omega + (N-1)\delta\omega/2$ is the average frequency of the wave packet. In view of the fact that $N\delta\omega = \Delta\omega$ is the total width of the frequencies of the wave packet, we can write (12.9) in the form

$$E(t) = E_0 \frac{\sin(\Delta\omega t/2)}{\sin[\Delta\omega t/(2N)]} \cos(\langle \omega \rangle t). \tag{12.10}$$

In most cases of practical importance, $N \gg 1$ and hence over a large number of periods of the argument $\Delta\omega t/2$ of the sine in the numerator of the formula, the argument of the sine in the denominator of the formula remains small ($\Delta\omega t/2N \ll 1$), so that we can put $\sin[\Delta\omega t/(2N)] \approx \Delta\omega t/(2N)$. In view of this, (12.10) can be written in the form

$$E(t) = E_0 N \frac{\sin(\Delta\omega t/2)}{(\Delta\omega t/2)} \cos(\langle \omega \rangle t). \tag{12.11}$$

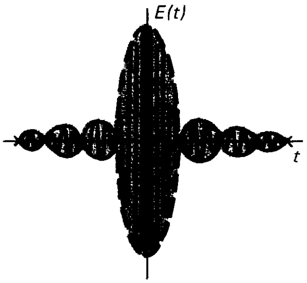


Fig. 53. A wave packet with equidistant frequencies.

The plot of this function is shown in Fig. 53. The dashed envelope represents the varying amplitude of oscillations in the wave packet, the fundamental frequency of these oscillations being $\langle\omega\rangle$. The energy of such a wave packet is concentrated over a comparatively small interval of frequencies. Hence wave packets are also called pulses. We shall be using both these terms according to circumstances. The maximum amplitude is observed at the point $t = 0$, when all oscillations are added in the same phase. In an interval Δt of time determined by the condition

$$\Delta\omega\Delta t/2 = \pi, \tag{12.12}$$

the oscillation amplitude becomes zero. This time is taken as a measure of the duration of the central pulse. Hence the frequency interval $\Delta\nu = 2\pi\Delta\omega$ between the component oscillations and the duration of the pulse are connected by the following relation:

$$\Delta\nu\Delta t \simeq 1, \tag{12.13}$$

where we have used the approximate equality to show that there exists a certain arbitrariness in the definition of the pulse duration. Such a relation was obtained earlier [see (8.56)] from the spectral analysis of a rectangular pulse. In view of the universal nature of relation (12.13), it is often called the frequency bandwidth theorem.

QUASI-PLANE WAVE. The plane wave represented by a formula of the type (2.50) can only be a wave that is spatially unbounded in all directions. The confinement of a wave in the direction of its propagation leads to its nonmonochromaticity which is characterized by the spectral width $\Delta\omega$ (see Sec. 9). The confinement of a wave to perpendicular directions leads to the emergence of a finite spectral width Δk_x and Δk_y of the wave numbers (the wave propagates along the Z-axis). In other words, a wave with a finite beam cross section cannot propagate strictly in one direction characterized by the vector k. There always exists a certain spread of the wave vector directions with respect to the mean direction. This phenomenon is called diffraction (see Chap. 6). Consequently, there are no plane waves with a finite cross section. However, if the spread in the directions of the wave vectors is not large, the wave can be assumed to be plane to a fairly high degree of accuracy and presented in the form (2.50), where k represents the mean wave vector of the wave. Such a wave is called a quasi-plane wave.

The mathematical criterion of a quasi-plane wave is ob-

!

Group velocity is the velocity of the maximum amplitude of a group of waves. This is the rate at which the energy of a wave packet propagates. If there is a dispersion, the group velocity differs from the phase velocity.

With the help of the substitution $\omega \to k$, $t \to z$, the relation between the emission linewidth and the pulse duration can be presented as a relation between the width of a wave packet and its spatial extent.

?

What does the frequency bandwidth theorem state? What is the width of the frequency band of a pulse represented by a δ-function?

tained by proceeding in the same way as for obtaining criterion (9.42) of a quasi-monochromatic wave. The wave can be presented in the form of a Fourier integral:

$$E(x, y, z, t) = \frac{1}{(2\pi)^4} \iiiint F(k_x, k_y, k_z, \omega)\, \mathrm{e}^{-i(\omega t - \quad \cdot \quad)}\, \mathrm{d}k_x\, \mathrm{d}k_y\, \mathrm{d}x_z\, \mathrm{d}\omega. \tag{12.14}$$

If the linear dimensions of the cross section of a wave propagating in the direction of the Z-axis are large in comparison with the wavelength, the amplitude F will be nonzero only in a narrow interval of wave numbers Δk_x, Δk_y near $k_x = 0$, $k_y = 0$. If

$$|2\Delta k_x| \ll k_z \text{ and } |2\Delta k_y| \ll k_z, \tag{12.15}$$

the wave is called a quasi-plane wave. It can be presented in the form (2.50) quite accurately if k is the mean wave vector. A quasi-plane wave can be considered plane only over a part of the wave front whose linear dimensions are smaller than the coherence width (see Secs. 27, 30). If, however, while considering some phenomenon we have to take into account the variation over a region of wave front larger than the coherence width, it cannot be considered as a plane wave or presented in the form (2.50).

SEC. 13. RANDOM LIGHT

Waves with random amplitudes and phases are analyzed.

SUPERPOSITION OF WAVES WITH RANDOM PHASES. In classical physics, an isolated stationary atom emits a train of waves with an exponentially decreasing amplitude of duration about 10^{-8} s. The width of the emission line is of the order of 10^8 Hz (see Sec. 9), and it has the Lorentz shape. As a result of interaction with other atoms during emission, collision broadening of the line takes place (see Sec. 10). At the instant t, the electric field strength of the wave at a certain point is given by the expression

$$E_i(t) = E_0 \mathrm{e}^{-i[\omega(t - t_i') - \varphi_i']}, \tag{13.1}$$

where t_i' is the instant at which the wave train is emitted after collision and φ_i' is the random initial phase of the radiation. The latter can be combined with the term $-\omega t_i'$ which depends on the distance between the atom and the point of observation, also a random quantity. Hence the electric field strength of the radiation field of the atom at a certain point can be presented

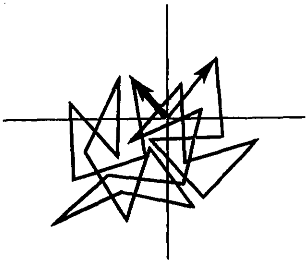


Fig. 54. Addition of instantaneous values of phase multipliers.

in the form

$$E_i(t) = E_0 e^{-i(\omega t - \varphi_i)}, \tag{13.2}$$

where $\varphi_i = \omega t_i' + \varphi_i'$ is the random phase. So far, we have not taken into consideration the Doppler effect which leads to a change in the frequency. It can be easily taken into account as an additional factor causing the broadening of the emission line. For the sake of simplicity, we assume that all the waves have the same amplitude and polarization. The total field strength due to all the radiators is equal to

$$E(t) = E_0 \sum_i e^{-i(\omega t - \varphi_i)} = E_0 e^{-i\omega t} \sum_i e^{i\varphi_i}, \tag{13.3}$$

where φ_i is the random phase of the wave train generated by the i-th atom. The series under the summation symbol can be presented as a vector sum of complex numbers. The modulus of each term in the series $\exp(-i\varphi_i)$ is equal to unity. Figure 54 shows the result of the addition of these terms at a certain moment of time t. Analytically, their sum can be written in the form

$$\sum_i e^{i\varphi_i} = a(t) e^{i\varphi(t)}, \tag{13.4}$$

where $a(t)$ is the amplitude of the sum of exponential terms and $\varphi(t)$ is its phase. Thus,

$$E(t) = E_0 a(t) e^{-i[\omega t - \varphi(t)]}, \tag{13.5}$$

where ω is the carrier frequency of the wave whose amplitude and phase have been modulated. The spectral composition of this wave is the same as for a collision-broadened line. This is so because significant changes in the phase and amplitude in (13.4) occur during a time τ_0 equal to the time between the perturbations. Hence the time parameters determining the collision broadening and the spectral width of the modulated wave (13.5) are of the same order. Consequently, the spectral composition of the radiation is basically the same.

RESOLUTION TIME. Each instrument measures just a certain value of a quantity averaged over a small interval of time called the resolution time. The resolution time of the best instruments for measuring the electric field strength is of the order of 10^{-9} s. Since the amplitude $a(t)$ and the phase $\varphi(t)$ in (13.5) change significantly over a time $\tau_0 \simeq 10^{-11}$ s, we conclude that these quantities can be treated as constants over tens or even hundreds of periods of oscillations. This means that if (13.5) is averaged over a period of oscillation or over several periods, $\langle E \rangle = 0$. Hence we can study experimentally

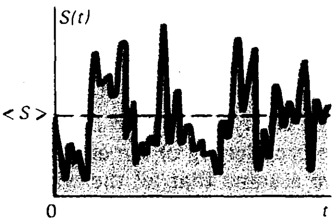


Fig. 55. Energy flux density fluctuations upon averaging over time intervals much shorter than τ_c.

not the mean values of the electric field strength (13.5), but rather the mean square values of the strength, i.e. the energy fluxes of the waves. The result of measurement of the wave energy flux in an experiment depends on the resolution time of the instrument.

AVERAGING OVER OSCILLATION PERIOD. Although there are no instruments having a resolution time equal to a single period of optical oscillations, it is expedient to analyze this question theoretically in order to discuss the results of averaging of the intensity over longer periods. Considering (3.4) and taking into account what has been said about the rate of variation of $a(t)$ and $\varphi(t)$ for the mean density of the energy flux of the waves whose electric field strength is given by (13.5), we obtain the formula

$$S = c(1/2)\varepsilon_0 E_0^2 a^2(t), \tag{13.6}$$

where the averaging is carried out over one period or over an interval of time that is much smaller than τ_0. To simplify the notation the averaging symbol is omitted from S and it is not stipulated that this is the mean density. The time variation of the flux density of the wave energy is determined by the quantity $a^2(t)$.

For such small averaging periods, the quantity $a^2(t)$ can be analyzed only theoretically by the methods of mathematical modelling. The main features of the time variation of $a^2(t)$ are clearly shown in Fig. 55. The dashed line shows the average values of $a^2(t)$ when averaging is carried out over time periods larger than τ_0.

Figure 55 shows that there are large fluctuations of the intensity $S(t)$, having the same order of magnitude as the mean value of the intensity. The time scale of fluctuations is determined by the value of τ_0. This is an alternative version of the statement that significant changes in $a(t)$ and $a^2(t)$ occur over a period of time τ_0.

EFFECT OF INCREASING TIME INTERVAL ON THE RESULT OF AVERAGING. If the interval of time over which the averaging is carried out is increased, the curve shown in Fig. 55 is smoothed out: the height of the peaks decreases and no sharp changes are observed. As the averaging time approaches τ_0 the fluctuation curve approaches the dashed line. As the averaging times become of the order of τ_0, the energy flux density of the waves does not change at all. This means that all the variations occur over an interval of time smaller than τ_0, and τ_0 is the scale of fluctuations. Further increases in the averaging time do not change the mean value. Thus, the time τ_0 is the characteristic time for the process under consideration.

?

Which quantity serves as the Fourier transform of a function describing a stationary stochastic process?
What is a spectral function? How can it lead to the autocorrelation function and the correlation interval?

COHERENCE TIME. The time τ_0 is the characteristic scale for random fluctuations in the phase and amplitude of the waves of a light beam with collision-broadened lines. For other mechanisms of line broadening and their combinations, there also exist similar characteristic times for random fluctuations in the phase and amplitude of a light beam. These characteristic times are called the coherence times τ_c and play a significant role in the interference phenomena (see Chap. 5).

COHERENCE LENGTH. The distance l_c covered by a light beam during the coherence time is called the coherence length:

$$\boxed{l_c = c\tau_c\,.} \tag{13.7}$$

For collision broadening, the coherence length is several thousand times more than the length of a light wave. For other mechanisms of line broadening the coherence length is also much more than the wavelength of light. The coherence length is the spatial scale of the fluctuations in the intensity of a light beam. From the time variation of fluctuations (Fig. 55), we can determine their variation in space if we take the t-axis as the Z-axis, the positive direction of the latter coinciding with the negative direction of the t-axis.

GAUSSIAN LIGHT. From the mathematical point of view, the random process of addition of phase multipliers (see Fig. 54) is a random walk problem which is considered, for example, in the theory of Brownian movement. In the present case, average length of a step is equal to unity. We denote by p the probability that after n steps, the final point in the random walk is determined by polar coordinates (a, φ). Then the probability for the final point to be on the area element $a\mathrm{d}a\mathrm{d}\varphi$ after n steps is given by the formula

$$\mathrm{d}P = pa\mathrm{d}a\mathrm{d}\varphi\,. \tag{13.8}$$

The random walk theory gives the following expression for p:

$$p = p(a) = (1/\pi n)\exp(-a^2/n) \tag{13.9}$$

with normalization

$$\int \mathrm{d}P = \int_0^\infty a\mathrm{d}a \int_0^{2\pi} p(a)\,\mathrm{d}\varphi = 2\pi \int_0^\infty ap(a)\,\mathrm{d}a = 1\,. \tag{13.10}$$

Since $a\mathrm{d}a\mathrm{d}\varphi$ is an area element, formula (13.8) indicates that the amplitudes $a(t)$ are distributed according to the Gaussian law (13.9). This distribution is valid for any source of random emission. Hence light from a random source is often called Gaussian light.

ENERGY FLUX DENSITY FLUCTUATIONS FOR RANDOM LIGHT. If averaging is carried out over a period longer than the coherence time τ_c ($\tau_c = \tau_0$ for collision broadening), we obtain

$$\langle a^2(t)\rangle = \langle|\sum_i e^{i\varphi_i(t)}|^2\rangle = n, \tag{13.11}$$

since the average value of the cross terms is equal to zero, $\langle|e^{i\varphi_i(t)}|^2\rangle = 1$, and n is the number of atoms whose radiation is being observed. In the case of averaging over a time interval much longer than the oscillation period, the flux density S [see (13.6)] represents the instantaneous value of the flux density of the wave energy, while averaging in accordance with (13.11) gives the average density. The average energy flux density of the waves obtained from (13.6) by taking into account (13.11) is:

$$\langle S\rangle = (1/2)\, c\varepsilon_0 E_0^2 n. \tag{13.12}$$

A comparison of (13.12) and (13.6) shows that

$$a^2(t) = Sn/\langle S\rangle. \tag{13.13}$$

Formula (13.8) then assumes the form

$$\mathrm{d}P = \frac{1}{\pi n}\exp\left(-\frac{a^2}{n}\right) a\,\mathrm{d}a\,\mathrm{d}\varphi = \frac{1}{2\pi}\frac{1}{\langle S\rangle}\exp\left(-\frac{S}{\langle S\rangle}\right)\mathrm{d}S\,\mathrm{d}\varphi. \tag{13.14}$$

Hence the distribution of the probability density S is given by the formula

$$p'(S) = (1/\langle S\rangle)\exp(-S/\langle S\rangle) \tag{13.15}$$

with normalization condition

$$\int_0^\infty p'(S)\,\mathrm{d}S = 1. \tag{13.16}$$

The energy flux density fluctuations can be calculated with the help of (13.14) and are characterized by the mean square deviation from the equilibrium value:

$$\sqrt{\langle(\Delta S)^2\rangle} = \sqrt{\langle(S - \langle S\rangle)^2\rangle} = \sqrt{\langle S^2\rangle - (\langle S\rangle)^2}, \tag{13.17}$$

where $\langle(S - \langle S\rangle)^2\rangle = \langle S^2\rangle - (\langle S\rangle)^2$. Since

$$\langle S^2\rangle = \frac{1}{\langle S\rangle}\int_0^\infty S^2 \exp\left(-\frac{S}{\langle S\rangle}\right)\mathrm{d}S = 2(\langle S\rangle)^2, \tag{13.18}$$

we obtain

$$\sqrt{(\Delta S)^2} = \langle S\rangle, \tag{13.19}$$

!

Random fluctuations of the phase and amplitude of a wave in a light beam are characterized by a certain minimum time interval τ_c, an averaging over which completely levels out the variation in the energy flux density. This characteristic is called the coherence time. Amplitudes of random radiation are distributed according to the Gauss law. In random light, the intensity fluctuation is equal in magnitude to the mean intensity.

?

Over what time intervals must the averaging be carried out in order to determine the intensity fluctuations and the average luminous flux intensity? Which circumstances are responsible for the Gaussian distribution of the wave amplitudes of random light? What is the physical meaning of the "instantaneous" value of the luminous flux intensity? Considering that the volume density of radiation is proportional to the number density of photons at a fixed frequency, find out, by using formula (13.19), the fluctuations in the number of photons falling on a detector over a fixed time interval. For what duration of this time interval can the statement concerning the fluctuation be verified experimentally?

i.e. the fluctuation in the energy flux density is equal to its mean value. Thus, as has been shown qualitatively in Fig. 55, the deviation of the energy flux density of waves from the mean in a light beam is quite large. It should be emphasized once again that the "instantaneous value" of the energy flux density in a light beam must be measured over a period of time much smaller than the coherence time, and the whole set of measurements must be made in a time exceeding the coherence time τ_c. While deriving relation (13.19), it was assumed that the instantaneous value of the energy flux density S is the value averaged over an oscillation period. In actual practice, such detectors do not exist. When the finite response time of the detector is taken into consideration, formula (13.19) is somewhat modified since the fluctuations are smoothed out to some extent.

POLARIZATION. Monochromatic electromagnetic waves (see Sec. 5) always have a certain polarization, i.e. are polarized.

In random light, which is a superposition of waves with random phases and amplitudes, the tip of the field vector of waves describes an irregular and highly chaotic curve in the plane perpendicular to the direction of the wave propagation. If these oscillations of the field vector do not have a preferred direction, the wave is called unpolarized. If, on the contrary, there exists a certain regularity in the oscillations of the vector, albeit not so clearly manifested as in a polarized wave, the wave is called partially polarized. The quantitative theory of partial polarization is based on the theory of coherence of mutually perpendicular components of the electric field strength of the wave. Hence it is expedient to describe this theory together with the coherence theory (see Sec. 30).

SEC. 14. FOURIER ANALYSIS OF RANDOM PROCESSES

It is shown how Fourier transformation can be applied to functions that do not satisfy the mathematical criteria for the applicability of Fourier transformation. The concepts and methods of the Fourier analysis of random processes are described.

POWER SPECTRUM (SPECTRAL FUNCTION). Most of the random processes are stable, i.e. their general nature does not change with time. This means that the functions describing these processes do not have a Fourier transform since they are nonabsolutely integrable (the function does not tend to zero as $t \to \pm \infty$). Consequently, we cannot apply the usual methods and concepts of spectral analysis to these functions. However, this is not even necessary since only the mean characteristics are of any interest in the random phenomena, and the phase relations between the harmonic components in the spectral expansion are insignificant. Also the functional dependence of random functions on time is not clearly known. Hence more appropriate concepts and quantities are used in the Fourier analysis of random processes.

To eliminate the problem of divergence of integrals, we determine $F_T(\omega)$, viz. the Fourier transform of a random stationary process $f(t)$, over a finite but large interval of time T:

$$F_T(\omega) = \int_{-T/2}^{T/2} f(t)\,\mathrm{e}^{-i\omega t}\,\mathrm{d}t. \tag{14.1}$$

It is shown by (9.22) that the frequency distribution of energy in the emission spectrum is characterized by the square of the modulus of spectral density, viz. $|F(\omega)|^2 = F(\omega)F^*(\omega)$. In a uniform process, the frequency distribution of energy does not vary with time, while the total emission energy is proportional to time. Hence

$$\int_{-\infty}^{\infty} |F_T(\omega)|^2\,\mathrm{d}\omega$$

increases in direct proportion to T and, consequently, $|F_T(\omega)| \propto \sqrt{T}$. Therefore, the quantity

$$\boxed{F_\mathrm{s}(\omega) = \lim_{T\to\infty} \frac{F_T(\omega)}{\sqrt{T}} = \lim_{T\to\infty} \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} f(t)\,\mathrm{e}^{-i\omega t}\,\mathrm{d}t} \tag{14.2}$$

is finite, independent of time, and can be used for characterizing the spectral properties of random processes. The square of the modulus of this quantity, viz.

$$\boxed{w_\mathrm{s}(\omega) = F_\mathrm{s}(\omega)F_\mathrm{s}^*(\omega)} \tag{14.3}$$

is called the power spectrum (spectral function). Using (14.2), we obtain

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-\infty}^{\infty} w_\mathrm{s}(\omega)\,\mathrm{d}\omega \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \mathrm{d}t f(t) \int_{-T/2}^{T/2} \mathrm{d}t' f(t') \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp[-i\omega(t-t')]\,\mathrm{d}\omega \\
&= \lim_{t\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t)\,\mathrm{d}t \int_{-T/2}^{T/2} f(t')\,\mathrm{d}t'\delta(t-t') \\
&= \lim_{t\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |f(t)|^2\,\mathrm{d}t = \langle |f(t)|^2 \rangle.
\end{aligned}$$

Thus, the spectral function is connected with the mean-square modulus of a random function through the relation

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} w_s(\omega)\,\mathrm{d}\omega = \langle |f|^2 \rangle_t. \tag{14.4}$$

The spectral function does not contain the phase characteristics of the process and, therefore, contains less information than the phase-amplitude spectrum. The spectral function describes the characteristics of the process that have been averaged over a large interval of time. Hence the small-scale characteristics of the process are not reflected in this spectrum.

AUTOCORRELATION FUNCTION. The most important characteristic of a random process is the interrelation between the values of the function describing this process at various points in space and time. The interrelation between the values of a function at the same point at different instants of time is described by the autocorrelation function. As an example, let us consider a stationary random quantity. We are interested in the relation between the value of this function at the instant t, i. e. $f_1 = f(t)$, and its value at the instant $t + \tau$, i. e. $f_2 = f(t + \tau)$, where τ is a certain fixed quantity. In other words, we must study the relation between the values of this function at instants of time separated by the interval τ. The instants of time themselves are arbitrary since the function is stationary. The only important thing is that they must be separated by a time interval τ. Due to the random nature of the function $f(t)$, the values $f(t)$ and $f(t+\tau)$ will have a random relation between themselves as a result of each measurement ("trial"). Hence the interrelation between the values of f at instants of time separated by the interval τ can be characterized only statistically.

The simplest statistical characteristic of a random quantity is its mean value

$$\langle f(t) \rangle_e = \langle f_1 \rangle, \quad \langle f(t+\tau) \rangle_e = \langle f_2 \rangle, \tag{14.5}$$

where the subscript "e" on the angle brackets indicates averaging over the ensemble or, in other words, the mathematical expectation over the set of possible values between $f(t)$ and $f(t + \tau)$. Apparently, this mean value does not depend on t for a stationary process, i. e. $\langle f_1 \rangle = \langle f_2 \rangle = \langle f \rangle$.

The interrelation between the values of f at instants of time separated by the interval τ is described by autocorrelation function which is defined as follows:

$$\Gamma_{11}(\tau) = \langle [f(t) - \langle f \rangle][f(t+\tau) - \langle f \rangle] \rangle_e. \tag{14.6}$$

If $p(f_1, f_2)$ is the probability density of the function $f(t)$ having the value f_1 and the function $f(t+\tau)$ having the value f_2, we can present (14.6) in the form

$$\Gamma_{11}(\tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (f_1 - \langle f \rangle)(f_2 - \langle f \rangle) p(f_1, f_2)\, \mathrm{d}f_1\, \mathrm{d}f_2. \quad (14.7\mathrm{a})$$

If the process is ergodic, the ensemble averaging can be replaced by time averaging; in this case, we can replace (14.7a) by

$$\begin{aligned}\Gamma_{11}(\tau) &= \langle [f(t) - \langle f \rangle][f(t+\tau) - \langle f \rangle] \rangle_t \\ &= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} [f(t) - \langle f \rangle][f(t+\tau) - \langle f \rangle]\, \mathrm{d}t, \end{aligned} \quad (14.7\mathrm{b})$$

where $\langle f \rangle$ is the time-averaged value of f.

In order to simplify the formulas, let us consider the function $f(t) - \langle f \rangle$ as a random quantity whose mean value is equal to zero. To avoid using additional symbols, we denote this function also by $f(t)$. For such a function, formula (14.7b) assumes the form

$$\Gamma_{11}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t) f(t+\tau)\, \mathrm{d}t, \quad (14.8)$$

i.e. coincides with (14.7b) for $\langle f \rangle = 0$. All the results following from formula (14.8) for $\langle f \rangle = 0$ can be easily extended to the general case (14.7b). Unless stated otherwise, we shall be using relation (14.8) henceforth, i.e. assume that $\langle f \rangle = 0$.

The autocorrelation function $\Gamma_{11}(\tau)$ for $\tau = 0$ is equal to the mean value of the square of $f(t)$, i.e. of a certain positive quantity. For small values of τ, it continues to have a positive nonzero value which decreases with increasing τ. The correlation is said to have a finite value in the domain of values of τ for which $\Gamma_{11}(\tau) \neq 0$. As τ increases, the correlation weakens, i.e. $\Gamma_{11}(\tau)$ is always smaller than $\Gamma_{11}(0)$. From (14.8) we obtain

$$\begin{aligned}\Gamma_{11}(-\tau) &= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t) f(t-\tau)\, \mathrm{d}t \\ &= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2-\tau}^{T/2-\tau} f(t'+\tau) f(t')\, \mathrm{d}t' = \Gamma_{11}(\tau), \end{aligned} \quad (14.9)$$

where the transition from the first to the second integral is made by using the substitution $t = t' + \tau$. Thus, an autocorrelation function is symmetric with respect to zero:

$$\Gamma_{11}(-\tau) = \Gamma_{11}(\tau). \quad (14.10)$$

If the values of the functions $f(t)$ and $f(t+\tau)$ are not interrelated, i.e. are independent of each other, the autocorrelation function is equal to zero. In particular, this function tends to zero as $\tau \to \infty$.

WIENER-KHINTCHINE THEOREM. The importance of the autocorrelation function is due, in particular, to its relation with the power spectrum which is established by the Wiener-Khintchine theorem: the spectral function is the Fourier transform of the autocorrelation function, while the autocorrelation function is the Fourier transform of the spectral function.

In order to prove this theorem, we transform the expression (14.3) with the help of (14.2):

$$w_s(\omega) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t') \mathrm{e}^{-i\omega t'} \mathrm{d}t' \int_{-T/2}^{T/2} f(t) \mathrm{e}^{i\omega t} \mathrm{d}t, \tag{14.11}$$

where we have considered that the function f is real ($f^* = f$). For further transformations, we observe that for stationary functions the values of these integrals are independent of the period of integration as long as the integration interval is T. Hence in the first integral we can replace the integration variable t' by $t+\tau$ and integrate with respect to $\tau\,(\mathrm{d}t' = \mathrm{d}\tau)$, leaving the limits of integration unchanged:

$$w_s(\omega) = \lim_{T \to \infty} \int_{-T/2}^{T/2} \mathrm{d}\tau \mathrm{e}^{-i\omega\tau} \frac{1}{T} \int_{-T/2}^{T/2} f(t) f(t+\tau)\, \mathrm{d}t. \tag{14.12}$$

In view of (14.8), the inner integral in (14.12) is equal to $\Gamma_{11}(\tau)$. Hence the above expression is transformed into the relation

$$w_s(\omega) = \int_{-\infty}^{\infty} \Gamma_{11}(\tau) \mathrm{e}^{-i\omega\tau} \mathrm{d}\tau, \tag{14.13}$$

which proves the first part of the Wiener-Khintchine theorem. The second part of the theorem is just the inverse transform of (14.13):

$$\Gamma_{11}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} w_s(\omega) \mathrm{e}^{i\omega\tau} \mathrm{d}\omega. \tag{14.14}$$

Taking the complex conjugate quantities on both sides of (14.2) and considering that the relation $f^* = f$ holds for the real function f, we conclude that $F_s^*(\omega) = F_s(-\omega)$ and hence (14.3) is an even function of ω, i.e. $w_s(\omega) = w_s(-\omega)$. The correlation function $\Gamma_{11}(\tau)$ is also an even function of τ in

!

The functions describing stationary stochastic processes do not have Fourier transforms. A special quantity serves as the Fourier transform for these functions. The square of the magnitude of this quantity is called the spectral function (power spectrum). The spectral function does not include phase characteristics of a random process. It contains the characterisics of a random process averaged over a large time interval. Hence it does not reflect the small-scale characteristics of the process. The spectral function is the Fourier transform of the autocorrelation function, and vice versa. By determining the spectral function experimentally, we can find the autocorrelation function. The correlation interval is equal in magnitude to the normalized spectral function at zero frequency.

accordance with (14.10). Hence, substituting $\exp(\pm i\omega\tau) = \cos\omega\tau \pm i\sin\omega\tau$ in the integrands of (14.13) and (14.14), we arrive at the conclusion that the integrals containing sine vanish and the relations assume a real form:

$$w_s(\omega) = 2\int_0^\infty \Gamma_{11}(\tau)\cos\omega\tau d\tau, \tag{14.15}$$

$$\Gamma_{11}(\tau) = \frac{1}{\pi}\int_0^\infty w_s(\omega)\cos\omega\tau d\tau. \tag{14.16}$$

The Wiener-Khintchine theorem can be used to determine the power spectrums if the autocorrelation function is known. The latter can be measured experimentally. Usually, it is a rapidly attenuating function and hence integral (14.15) can be solved without any difficulty. Thus the spectral function can be also determined experimentally.

CORRELATION INTERVAL. It is obvious from (14.8) that

$$\Gamma_{11}(0) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} f^2(t)\,dt = \langle f^2\rangle_t. \tag{14.17}$$

Hence instead of the correlation function $\Gamma_{11}(\tau)$ it is convenient to use the dimensionless function

$$\gamma_{11}(\tau) = \Gamma_{11}(\tau)/\Gamma_{11}(0) = \Gamma_{11}(\tau)/\langle f^2\rangle, \tag{14.18}$$

called the normalized correlation function. It can be seen from (14.18) that $\gamma_{11}(0) = 1$. The measure of correlation is the quantity

$$\tau_{\text{cor}} = \int_{-\infty}^{\infty}\gamma_{11}(\tau)\,d\tau = 2\int_0^\infty \gamma_{11}(\tau)\,d\tau \tag{14.19}$$

called the correlation interval.

RELATION BETWEEN THE CORRELATION INTERVAL AND THE NORMALIZED SPECTRAL FUNCTION. In analogy with the normalized correlation function, we introduce the normalized spectral function

$$w_n(\omega) = w_s(\omega)/\langle f^2\rangle = w_s(\omega)/\Gamma_{11}(0). \tag{14.20}$$

The correlation interval and the normalized spectral function are connected through an important relation. In order to derive this relation, we express $\gamma_{11}(\tau)$ in (14.19) in terms of the normalized spectral function through formula (14.14):

$$\tau_{\text{cor}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w_{\text{n}}(\omega)\, e^{i\omega\tau}\, d\omega d\tau = \int_{-\infty}^{\infty} w_{\text{n}}(\omega)\, \delta(\omega)\, d\omega = w_{\text{n}}(0). \tag{14.21}$$

Thus, the correlation interval is equal to the normalized spectral function at zero frequency.

PROBLEMS

2.1. Find the coherence length for the radiation from a ruby laser ($\lambda = 693.6$ nm) if the width of the emission line in wavelengths is equal to $\Delta\lambda = 1.6 \times 10^{-17}$ m.

2.2. Write the Fourier series for a periodic function with a period T, which is equal to 2 in the interval $(-T/4, +T/4)$, and zero outside this interval up to its boundaries.

2.3. Write the Fourier series for a periodic function with a period T, having the value 2 in the interval $(0, T/2)$ and zero in the interval $(T/2, T)$.

2.4. Write the Fourier series for a periodic function with a period T, which is described by the formula $f(t) = t/T$ in the interval $(0, T)$.

2.5. Write the Fourier series for a periodic function with a period T, which is described by the following formula

$$f(t) = \begin{cases} -2t/T & \text{for } -T/2 < t < 0, \\ 2t/T & \text{for } 0 < t < T/2. \end{cases}$$

2.6. Find the Fourier transform of the function

$f(t) = \exp(-\alpha t^2)\cos\omega_0 t\,(\alpha > 0)$.

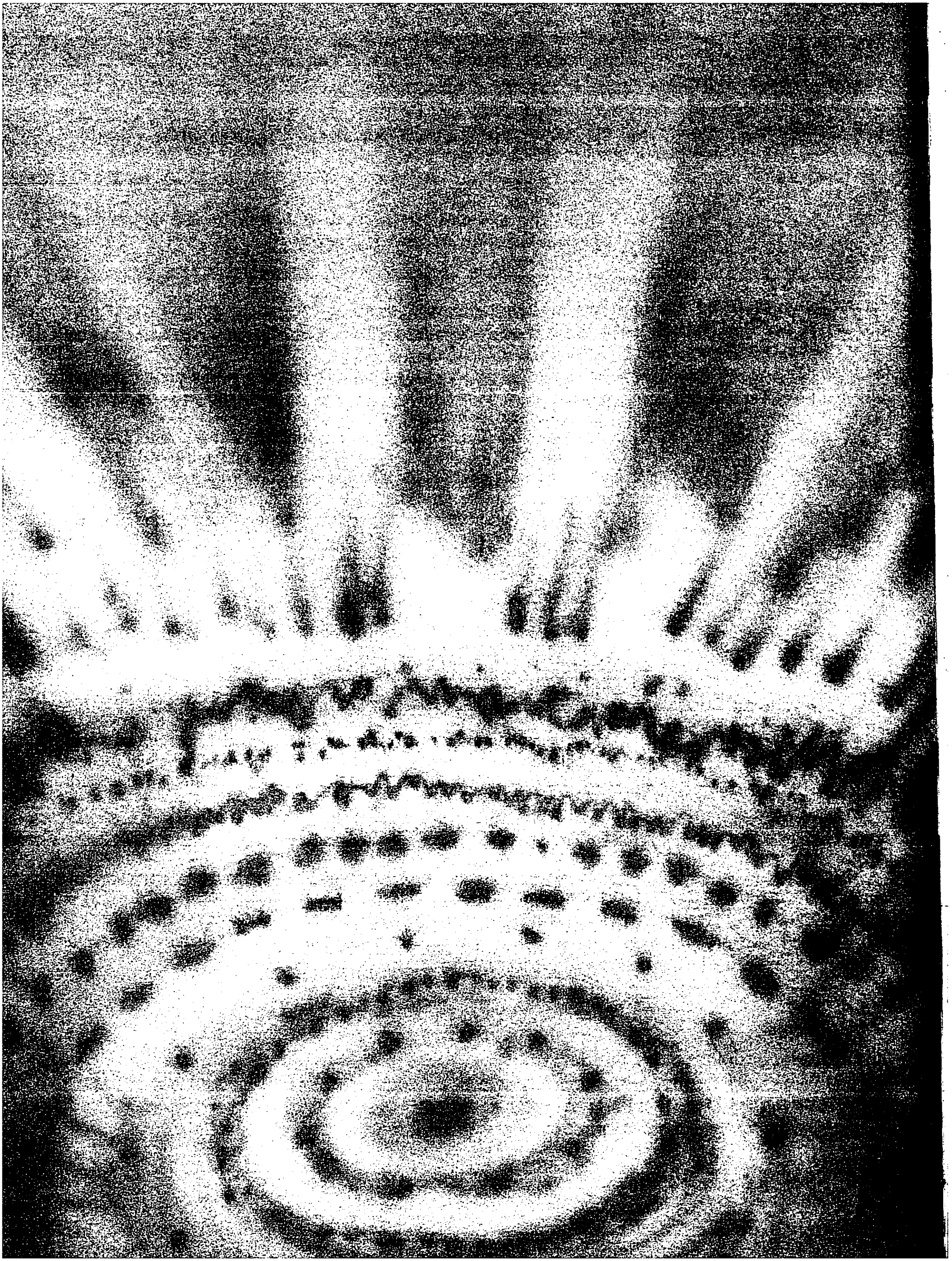
2.7. Find the Fourier transform of the function

$f(z) = 4\exp(-z^2/a^2)$.

2.8. Make calculations to prove that a line formed by combining two Lorentz lines is a Lorentz line with a width $\gamma = \gamma_1 + \gamma_2$, while the line formed by combining two Gaussian lines is a Gaussian line with a width $\Delta = \sqrt{\Delta_1^2 + \Delta_2^2}$.

ANSWERS

2.1. 3×10^4 m. 2.2. $1 + (4/\pi)[\cos\omega t - 1/3\cos 3\omega t + 1/5\cos 5\omega t - \ldots]$. 2.3. $1 + (4/\pi)(\sin\omega t + 1/3\sin 3\omega t + 1/5\sin 5\omega t + \ldots)$. 2.4. $f(t) = 1/2 - (1/\pi)(\sin\omega t + (1/2)\sin 2\omega t + \ldots)$. 2.5. $f(t) = 1/2 - (4/\pi^2) \times [\cos\omega t + (1/3^2)\cos 3\omega t + (1/5^2)\cos 5\omega t + \ldots]$. 2.6. $\sqrt{\pi/\alpha}\exp \times [-(\omega - \omega_0)^2/(4\alpha)]$. 2.7. $aA\sqrt{\pi}\exp - (k^2a^2/2)$.



Chapter 3

Propagation of Light in Isotropic Media

Basic idea:

The properties of a medium are described in terms of scalar quantities, viz. permittivity, permeability and conductivity. The behaviour of light at the interface between different media is determined by the boundary conditions.

SEC. 15. PROPAGATION OF LIGHT IN DIELECTRICS

The microscopic mechanism of emergence of dispersion and its manifestations are described. The propagation of pulses in a dispersive medium is studied.

MONOCHROMATIC WAVES. The permittivity ε of a homogeneous isotropic dielectric is independent of coordinates. It is also assumed to be independent of time. In this case, Maxwell's equations are similar to Eqs. (2.1)-(2.5), but ε_0 is replaced by ε, i.e. only the first of equations (2.5) is different:

$$\mathbf{D} = \varepsilon \mathbf{E}. \tag{15.1}$$

Hence all the subsequent results of Sec. 2 for electromagnetic waves in vacuum are also valid for a dielectric, the only difference being that ε_0 is replaced by ε. This causes only a change in the wave velocity. Substituting ε_0 for ε, we obtain from (2.8) and (2.9) the following expression for the velocity of electromagnetic waves in a dielectric:

$$v = 1/\sqrt{\varepsilon\mu_0} = 1/\sqrt{\varepsilon_r\varepsilon_0\mu_0} = c/\sqrt{\varepsilon_r} = c/n, \tag{15.2}$$

where $\varepsilon_r = \varepsilon/\varepsilon_0$ is the relative permittivity, $1/\sqrt{\varepsilon_0\mu_0} = c$ is the velocity of light and $n = \sqrt{\varepsilon_r}$ is the refractive index of the dielectric relative to vacuum, or simply the refractive index.

Electromagnetic waves in a dielectric are analogous to waves in vacuum (see Sec. 2). There is no need to repeat all that was stated in Sec. 2. Suffice it to say that the wavelength λ is connected with the frequency ω through relation (2.39) in which c has been replaced by v:

$$\lambda = vT = 2\pi v/\omega, \tag{15.3}$$

while the wave number is defined by the relation

$$k = 2\pi/\lambda = \omega/v. \tag{15.4}$$

The expression for the wave vector is also modified accordingly.

The structure of a plane electromagnetic wave in a dielectric described by Eqs. (2.53)-(2.56) with the replacement of ε_0 by ε is similar to that in vacuum. Relation (2.57) assumes the form

$$E = vB. \tag{15.5}$$

The energy flux density of the waves in a dielectric is given by formula (3.1). Instead of (3.2), we obtain the following expression for $|\mathbf{S}|$:

$$S = |\mathbf{S}| = |\mathbf{E}|\,|\mathbf{H}| = EB/\mu_0 = E^2/(\mu_0 v), \tag{15.6}$$

where we have taken into account Eq. (15.5). Since $1/\mu_0 = \varepsilon v^2$ in a dielectric [see (15.2)], the above expression assumes the form

$$S = v\varepsilon E^2. \tag{15.7}$$

The volume energy density of electromagnetic field in a dielectric is expressed by the formula

$$w = (1/2)(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H}) = \varepsilon E^2, \tag{15.8}$$

where $\mathbf{B}\cdot\mathbf{H} = E^2/(\mu_0 v^2) = \varepsilon E^2$. Hence the interpretation of relation (15.7) becomes quite clear if we write it in the form

$$\boxed{S = vw.} \tag{15.9}$$

Instead of (3.4) we obtain for the time-averaged value of the energy flux density of waves

$$\langle S\rangle_t = v\varepsilon E_0^2/2, \tag{15.10}$$

where E_0 is the electric field strength of the wave.

All that has been stated in Sec. 3 about the distribution of the energy flux density over the beam cross section or about the energy flow rate and other related questions remains valid for a homogeneous dielectric and need not be repeated.

DISPERSION. The velocity of electromagnetic waves in a dielectric depends on frequency. This phenomenon is called dispersion. The effect of dispersion is manifested only in the propagation of nonmonochromatic waves, since various frequencies that make up a wave propagate with different frequencies.

Dispersion is the consequence of the frequency dependence of atomic polarization. The motion of an electron in a wave electric field of strength

$$E = E_0 e^{-i\omega t} \tag{15.11}$$

is described by Eq. (9.32). The solution of this equation is given by formula (9.36). The dipole moment of the atom whose electron moves from its equilibrium position ($x = 0$) to the point x is given by

$$p = ex = e^2E/[m(\omega_0^2 - \omega^2 - i\gamma\omega)]. \tag{15.12}$$

Hence the frequency-dependent and time-varying polarization is

$$P = Np = e^2NE/[m(\omega_0^2 - \omega^2 - i\gamma\omega)], \tag{15.13a}$$

where N is the number density of electrons with the natural frequency ω_0. Equation (15.13a) can be written in the form

$$P = \varepsilon_0 \varkappa^{(1)} E, \tag{15.13b}$$

where

$$\varkappa^{(1)} = e^2N/[m\varepsilon_0(\omega_0^2 - \omega^2 - i\gamma\omega)] \tag{15.13c}$$

is the (linear) complex dielectric susceptibility. If there are electrons with another natural frequency of oscillations, we must add a corresponding term to the right-hand side of (15.13). Taking into account the relation

$$D = \varepsilon_0 E + P \tag{15.14}$$

and considering that $D = \varepsilon_\omega E$, while P is given by formula (15.13a), we obtain from (15.14) after cancelling out the common factor E

$$\varepsilon_\omega = \varepsilon_0 + e^2N/[m(\omega_0^2 - \omega^2 - i\gamma\omega], \tag{15.15}$$

where ε_ω is the frequency-dependent permittivity. From (15.15) it follows that

$$\varepsilon_{r\omega} = \varepsilon_\omega/\varepsilon_0 = 1 + e^2N/[m\varepsilon_0(\omega_0^2 - \omega^2 - i\gamma\omega)], \tag{15.16}$$

where $\varepsilon_{r\omega}$ is the frequency-dependent relative permittivity. In accordance with (15.2), this means that the refractive index $n'_\omega = \sqrt{\varepsilon_{r\omega}}$, and hence the velocity of electromagnetic waves depend upon frequency. This explains the mechanism of emergence of dispersion.

In accordance with (15.16), we can write the following expression for the refractive index:

$$n'^2_\omega = \varepsilon_{r\omega} = 1 + e^2N/[m\varepsilon_0(\omega_0^2 - \omega^2 - i\gamma\omega)]. \tag{15.17}$$

It follows from this relation that n'_ω is a complex quantity. We represent it in the form

$$\boxed{n'_\omega = n_\omega + i\xi_\omega,} \tag{15.18}$$

where n_ω and ξ_ω are the real and imaginary parts of n'_ω. Substituting (15.18) into (15.17) and equating the real and imaginary parts of the obtained equality, we get two equations:

$$n_\omega^2 - \xi_\omega^2 = 1 + \frac{e^2 N}{\varepsilon_0 m} \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}, \tag{15.19}$$

$$2n_\omega\xi_\omega = \frac{e^2 N}{\varepsilon_0 m} \frac{\omega\gamma}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}. \tag{15.20}$$

Since γ is very small [see (9.7)] we find that in the optical range

$$\gamma^2\omega^2 \ll (\omega_0^2 - \omega^2)^2 \tag{15.21}$$

with the exception of the frequencies $\omega \approx \omega_0$. Hence it can be assumed that $\xi_\omega = 0$. In this case, formula (15.19) assumes the form

$$n_\omega^2 = 1 + e^2 N/[\varepsilon_0 m(\omega_0^2 - \omega^2)]. \tag{15.22}$$

If in addition to electrons with natural oscillation frequency ω_0 there also exist electrons with other natural oscillation frequencies, the terms corresponding to these frequencies must be added to the right-hand side of (15.22). Denoting by N_i the number density of electrons with natural oscillation frequency ω_{0i} in a dielectric, we obtain

$$n_\omega^2 = 1 + \frac{e^2}{\varepsilon_0 m} \sum_i \frac{N_i}{\omega_{0i}^2 - \omega^2}. \tag{15.23}$$

However, dispersion is caused not only by the electron oscillations. The oscillations of ions must also be taken into consideration. In view of the large mass of ions, their natural oscillation frequencies are much lower than for electrons. The ion oscillations are taken into account through additional terms on the right-hand side of (15.23). For example, the presence of singly charged ions with natural frequency ω_{0j} and number density N_j leads to the appearance of an additional term $e^2 N_j/[\varepsilon_0 M_j(\omega_{0j}^2 - \omega^2)]$, where M_j is the mass of the ion. Such terms are not included in (15.23) so as to avoid complication of the formula. The natural frequencies of ions lie in the far infrared region of the spectrum and do not significantly affect the shape of the dispersion curve in the visible range. However, they play a significant role in explaining why the static permittivity differs from permittivity in the visible range of frequencies. Formula (15.23) also requires a further refinement. The electrons subjected to an electromagnetic field do not oscillate identically, and some of them even

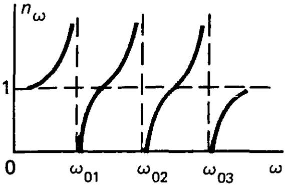


Fig. 56. Normal dispersion.

do not oscillate at all. Hence not all of the N_j electrons contribute to the dispersion, but only a fraction of these electrons is involved. Hence the modified formula in (15.23) must contain the quantity $N_i f_i$ instead of N_i, where $f_i \ll 1$ is called the oscillator strength. Formula (15.23) gives a satisfactory agreement with the experiment only if the oscillator strength is taken into account. The oscillator strength can be theoretically calculated only within the framework of quantum theory. The oscillator strength is not explicitly considered in (15.23) or in subsequent formulas.

Water can serve as an illustration of the contribution of ions to dispersion. For water, $\sqrt{\varepsilon_r} = n_\omega = 1.33$ for optical frequencies, while the static value $\sqrt{\varepsilon_r} \simeq 9$. This anomaly is explained by the ion oscillations. If these oscillations are taken into account, the right-hand side of (15.23) acquires additional terms of the type $e^2 N_j f_j / (\varepsilon_0 M_j \omega_{0j}^2)$ for $\omega \to 0$. These terms have large numerical values since the natural frequencies ω_{0j} of the ion oscillations are low. Near the minimum frequency ω_{0j}, the dispersion curve rises, while it drops as $\omega \to 0$, intersecting the axis $\omega = 0$ at $\varepsilon_r \simeq 81$, which explains the large static value of relative permittivity of water.

NORMAL DISPERSION. If the refractive index is close to unity as, for example, for rarefied gases, then $(n_\omega^2 - 1) = (n_\omega - 1)(n_\omega + 1) \simeq 2(n_\omega - 1)$ and formula (15.23) is simplified:

$$n_\omega = 1 + \frac{e^2}{2\varepsilon_0 m} \sum_i \frac{N_i}{\omega_{0i}^2 - \omega^2}. \tag{15.24}$$

The graph of the dependence of n_ω on frequency ω is called the dispersion curve (Fig. 56). If the refractive index increases with frequency, the dispersion is called normal. Figure 56 shows regions of the dispersion curve corresponding to normal dispersion. It can be stated that normal dispersion is observed in the entire transparent region.

For low frequencies $\omega \ll \omega_{0i}$, formula (15.24) gives the static value of the refractive index

$$n = 1 + \frac{e^2}{2\varepsilon_0 m} \sum_i \frac{N_i}{\omega_{0i}^2}, \tag{15.25}$$

which may differ significantly from the refractive index for optical frequencies.

For high frequencies ($\omega \gg \omega_{0i}$), the refractive index tends to

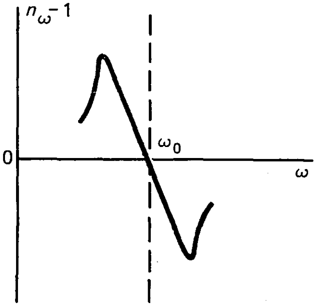


Fig. 57. Anomalous dispersion.

!

Normal dispersion is observed throughout the transparent region, while anomalous dispersion takes place in the absorption regions. The shape of a pulse varies in the course of its propagation through a dispersive medium. The leading front of the pulse in a dispersive medium moves with the velocity of light in vacuum. Only the rest of the pulse moves with a velocity which is characteristic of the medium. The primary wave in a medium is replaced over a certain path length by a secondary wave whose velocity of propagation is a characteristic property of the medium. This path length is equal to 0.5 mm for air, while for the interstellar space its value is about two light years.

unity, remaining less than this value all the time, since in this case formula (15.24) assumes the form

$$n_\omega = 1 - \frac{e^2}{2\varepsilon_0 m}\frac{1}{\omega^2}\sum_i N_i. \tag{15.26}$$

Consequently, a dielectric is an optically less dense material than vacuum for shortwave radiation. In particular, total reflection may be observed here from the surface of the dielectric (see Sec. 17), as is the case of, say, X-rays. It is obvious from (15.26) that at very high frequencies the nature of the electron bonds in the atom does not play any role, and the refractive index depends only on the total number of oscillating electrons per unit volume.

ANOMALOUS DISPERSION. The dispersion curve shown in Fig. 56 is plotted in accordance with formula (15.23) which is obtained from (15.17) without taking into account the damping of oscillations ($\gamma = 0$). This makes the refractive index n_ω turn to infinity at the points $\omega = \omega_{0i}$. If damping is taken into consideration ($\gamma \neq 0$), the dispersion curve becomes continuous in the neighbourhood of the points ω_{0i} and does not turn to infinity at the points $\omega = \omega_{0i}$. Let us express the refractive index in the form (15.18). If $|n'_\omega|$ differs slightly from unity, we obtain from (15.17)

$$n'_\omega = n_\omega + i\xi_\omega = \sqrt{\varepsilon_{r\omega}} \simeq 1 + e^2 N/[2\varepsilon_0 m(\omega_0^2 - \omega^2 - i\gamma\omega)]. \tag{15.27}$$

Separating the real and imaginary parts, we obtain

$$n_\omega = 1 + \frac{e^2}{2\varepsilon_0 m} N \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}, \tag{15.28}$$

$$\xi_\omega = \frac{e^2}{2\varepsilon_0 m} N \frac{\gamma\omega}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}. \tag{15.29}$$

The dispersion curve near the resonance frequency $\omega = \omega_0$ has the form shown in Fig. 57. Near the resonance frequency, the refractive index decreases with increasing frequency. This phenomenon is called anomalous dispersion.

SCATTERING OF LIGHT. The electric vector of a plane wave propagating in the direction of the Z-axis can be presented in the form

$$E(z, t) = E_0 e^{-i(\omega t - kz)}, \tag{15.30}$$

where

$$k = \omega/v = \omega\sqrt{\varepsilon_{r\omega}}/c. \tag{15.31}$$

Replacing $\sqrt{\varepsilon_{r\omega}}$ in (15.31) by its expression from (15.27), we obtain the relation

$$k = \omega n_\omega/c + i\omega\xi_\omega/c; \tag{15.32}$$

substituting this relation into (15.30), we get

$$E(z, t) = E_0 e^{-\omega\xi_\omega z/c} e^{-i(\omega t - \omega n_\omega z/c)}. \tag{15.33}$$

This formula shows that the imaginary part ξ_ω of the refractive index describes the damping of a plane wave in a dielectric. The damping is due to the fact that the wave energy is consumed on exciting forced oscillations in electrons. In turn, the oscillating electrons emit electromagnetic waves of the same frequency in all directions. Thus, the net result of this chain of processes is that the electromagnetic waves are scattered upon passing through a dielectric. This scattering is insignificant in view of the fact that γ is small.

PROPAGATION OF A WAVE PACKET. In the general case, a wave packet is a superposition of a group of waves whose frequencies and wave numbers are confined to quite narrow intervals. A wave packet can be presented in the form

$$E(z, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(k) e^{-i(\omega t - kz)} dk, \tag{15.34}$$

where the amplitude $F(k)$ is nonzero only in a narrow interval of values of k near k_0. For the independent variable, we have chosen the wave number k, and the frequency is a function of k, i.e. $\omega = \omega(k)$. In (15.34), it is assumed that all the waves in the group have the same polarization. The amplitude $F(k)$ characterizes the distribution of the waves over wave numbers, and hence over frequencies. It can be obtained from (15.34) for $t = 0$ with the help of Fourier transformation:

$$F(k) = \int_{-\infty}^{\infty} E(z, 0) e^{-ikz} dz. \tag{15.35}$$

This means that $F(k)$ is the Fourier transform of $E(z, 0)$.

If the frequency distribution of amplitudes of a wave packet has a sharp peak near the wave number k_0, the function $\omega(k)$ can be expanded near k_0 into a Taylor series in deviations of k from k_0, i.e. in $(k - k_0)$. Confining to the first term, we obtain

$$\omega(k) \approx \omega_0 + (k - k_0)\, d\omega_0/dk_0, \tag{15.36}$$

where $\omega_0 = \omega(k_0)$, $d\omega_0/dk_0 = (d\omega/dk)_{k=k_0}$. Formula (15.34) then assumes the form

$$E(z, t) \approx \frac{1}{2\pi} \exp\left[-i\left(\omega_0 - k_0 \frac{d\omega_0}{dk_0}\right) t\right] \int_{-\infty}^{\infty} F(k) \exp \times \left[ik\left(z - \frac{d\omega_0}{dk_0} t\right)\right] dk, \tag{15.37}$$

where the approximate equality emphasizes that only the first-order term in the Taylor series for $\omega(k)$ have been used in the calculations. Formula (15.34) shows that the integral in (15.37) is equal to $E[z-(\mathrm{d}\omega_0/\mathrm{d}k_0)t, 0]$ and therefore [see (15.37)]

$$E(z,t) \approx E[z-(\mathrm{d}\omega_0/\mathrm{d}k_0)t, 0]\exp\{-i[\omega_0 - k_0(\mathrm{d}\omega_0/\mathrm{d}k_0)]t\}. \tag{15.38}$$

Hence it can be seen that the square of the pulse amplitude is

$$|E|^2 = |E[z-(\mathrm{d}\omega_0/\mathrm{d}k_0)t, 0]|^2. \tag{15.39}$$

This means that to a first approximation, the pulse moves without any change in its shape with a group velocity

$$v_g = \mathrm{d}\omega_0/\mathrm{d}k_0 \tag{15.40}$$

[see (12.7)]. The inclusion of higher-order terms in the Taylor expansion for $\omega(k)$ indicates that the pulse shape does change during propagation.

All the formulas characterizing the motion of a pulse were obtained in approximation (15.36). In this case, it may so happen that the group velocity in the region of anomalous dispersion is higher than the velocity of light. This is not possible since energy is transmitted at the group velocity, and the transport of mass and energy at velocities higher than the velocity of light is inadmissible. This controversy is removed if we note that in the region of anomalous dispersion, approximation (15.36) is insufficient and we must take into consideration the higher-order terms in the expansion.

Dispersion appears as a result of the interference of primary and secondary waves. Hence the leading front of light pulse propagates in a medium with the velocity of light in vacuum, since the secondary waves cannot catch up with it. This part of the pulse is the first to arrive and has a small amplitude. This is followed by the second part of the pulse which has a higher amplitude and longer duration. The main signal arrives after this. Obviously, the signal velocity is not a well-defined concept, since we could consider the part of the pulse arriving first at the point of reception as the signal. Usually, when we speak of the velocity of a signal, we mean the group velocity at a frequency corresponding to the maximum amplitude in the signal. However, if a very sensitive detector were used, we could interpret the signal velocity as the velocity of the predecessors of the main signal. In this case, the velocity of the signal can be as close to the velocity of light in vacuum as possible, although the signal is still propagating in the medium.

REPLACEMENT OF A LIGHT WAVE IN A MEDIUM. A light wave entering a dielectric medium induces oscillations of dipoles which emit a secondary wave. The secondary wave propagates in the medium with a velocity that is characteristic of the medium; the primary wave continues to move in the medium at its initial velocity. The primary and the secondary waves are combined together. At the instant of its emission, the secondary wave has the same phase of oscillations as the primary wave. However, the difference in the phase velocities of propagation of the two waves in the medium causes a phase difference between them. As a result, the secondary wave suppresses the primary wave. Thus, the primary wave is replaced by the secondary wave which propagates with a velocity that is characteristic of the medium.

If the refractive index of the medium is denoted by n, a phase difference $\Delta\varphi = \omega|n-1|x/c$ appears between the waves over a distance x. For the substitution length we take the distance over which $\Delta\varphi \simeq 1$. Hence the substitution length is equal to

$$L = c/(\omega|n-1|) = \lambda/(2\pi|n-1|). \tag{15.41}$$

For transparent bodies like glass ($n \simeq 1.5$-2), the substitution length in the optical range is $L \simeq (2\text{-}1) \times 10^{-7}$ m, while for air and gases $L \simeq 0.5$ mm. In the interstellar space, the value of $(n-1)$ is extremely small and hence the substitution length is quite large. Calculations show that the substitution length for light in the interstellar space is about two light years.

DISPERSION OF LIGHT IN INTERSTELLAR SPACE. In the interstellar space, the dispersion of light is extremely small. The upper limit of dispersion can be estimated by observing pulsars. The brightness of pulsars varies periodically, the pulsation periods being several thousandths of a second in some cases. In the case of dispersion, the waves of different lengths have different times of propagation from the instant of their emission to the moment of observation. Hence waves of different frequencies, emitted by a pulsar at different instants of time, are observed simultaneously. This means that in the case of dispersion the curve showing the variation of the total brightness must be "smeared". The absence of such a "smearing" leads to an estimate for the upper limit of dispersion. Let $v(\omega)$ be the velocity of waves of frequency ω, Δt the duration of a radiation pulse, and L the distance between the pulsar and the observer. The absence of "smearing" means that the difference in the time of propagation of waves of different wavelengths must be less than the pulse duration:

$$\left|\frac{L}{v(\omega_1)} - \frac{L}{v(\omega_2)}\right| \leqslant \Delta t. \tag{15.42}$$

This inequality can be written in the following form:

$$\left|\frac{v(\omega_2) - v(\omega_1)}{c}\right| \leqslant \frac{c\Delta t}{L}, \tag{15.43}$$

where $c^2 \approx v(\omega_1)v(\omega_2)$. The following data are known for the Crab nebula: $L \simeq 6 \times 10^3$ light years, and $\Delta t = 3 \times 10^{-3}$ s. Hence the results of observation of this pulsar lead to the statement that $\Delta v/c < 10^{-14}$. Since the interval of observable frequencies extends from 4×10^8 to 3×10^{20} Hz, we can conclude with a high degree of accuracy that there is practically no dispersion.

COLOUR OF BODIES. The existence of resonance frequencies results in the coloration of bodies due to selective absorption of light near the resonance frequencies ω_{0i}. This may occur in the bulk of a body or in its surface layer if in the process of scattering light penetrates the bulk material and undergoes selective absorption.

If the frequency ω_{01} (see Fig. 56) lies in the ultraviolet region, there is no selective absorption in the visible part of the spectrum, and the material appears almost colourless and transparent. For example, ω_{01} for glass lies in the near-ultraviolet region close to the boundary of visible spectrum. Hence it is transparent but has a fairly high refractive index (about 1.5) and strongly absorbs ultraviolet rays. If the natural frequencies ω_{0i} lie in the visible part of the spectrum, the corresponding parts of the visible spectrum are absorbed and the material appears to have an additional colour. For example, the colour of the violet ink in a pot is due to the absorption of the yellowish part of the spectrum in the bulk when white light passes through the pot. The violet ink has also a violet colour on a sheet of paper. This is so because the fibres in the surface layer of the paper are impregnated with ink. The reflection (scattering) on light takes place over a surface layer of paper of a certain thickness, where the fibres impregnated with ink absorb the yellowish part of the spectrum. However, if a drop of ink falls on a surface that does not absorb it (for example, glass surface), the dried ink-stain will be yellowish in colour. The appearance of this colour is not due to selective absorption of light but due to selective reflection. The difference in the colours of materials upon selective absorption of light and upon reflection is manifested very effectively in gold. A chunk of gold has a reddish-yellow colour. However, if a very thin foil of gold is prepared, it has a deep blue colour when viewed through a clearance. In the former case, the colour of gold is due to selective reflection, while in the latter case it is due to selective absorption.

?

In which range do the natural oscillation frequencies of ions lie?

What is the physical meaning of the statement that for a very short-wave radiation a dielectric is an optically less dense medium than vacuum?

Is it possible in principle to transmit a signal through a medium at a velocity larger than the group velocity of light?

When light is reflected from a surface (see Secs. 16 and 20), the most strongly reflected regions of the spectrum are those which are absorbed most strongly upon passing through the bulk of the material. Hence the colour of a substance due to selective reflection is in addition to the colour of the same substance due to selective absorption.

Example 15.1. Express the group velocity (15.40) as a function of the refractive index and wavelength. Find the group velocity in water for $\lambda = 656.3$ nm if the refractive index of water at 20 °C is 1.3311 for this wavelength, and 1.3314 for $\lambda = 643.8$ nm.

Since $\omega = 2\pi c/(n\lambda)$, $k = 2\pi/\lambda$, we obtain from (15.40)

$$v_g = \frac{\mathrm{d}\,[2\pi c/(n\lambda)]}{\mathrm{d}\,(2\pi/\lambda)} = c\frac{(n\mathrm{d}\lambda + \lambda\mathrm{d}n)/(n^2\lambda^2)}{\mathrm{d}\lambda/\lambda^2} = \frac{c}{n}\left(1 + \frac{\lambda\,\mathrm{d}n}{n\,\mathrm{d}\lambda}\right). \tag{15.44}$$

Substituting numerical values into this equation, we obtain

$$v_g = \frac{3\times 10^8}{1.3311}\left(1 + \frac{656.3}{1.3311}\,\frac{0.0003}{12.5}\right)\ \text{m/s} = 2.28\times 10^8\ \text{m/s}.$$

SEC. 16. REFLECTION AND REFRACTION OF LIGHT AT THE INTERFACE BETWEEN TWO DIELECTRICS. FRESNEL'S FORMULAS

It is shown that the boundary conditions for the electric and magnetic field vectors of waves completely determine the laws of reflection and refraction.

BOUNDARY CONDITIONS. Dielectrics have different values of permittivity. The behaviour of a wave at a boundary is completely determined by the boundary conditions for the field vectors of the waves. In the absence of free charges and conduction currents, these conditions have the form

$$D_{2n} = D_{1n},\quad B_{2n} = B_{1n}, \tag{16.1a}$$

$$E_{2\tau} = E_{1\tau},\quad H_{2\tau} = H_{1\tau}, \tag{16.1b}$$

where the subscripts n and τ denote the normal and tangential components of the vector respectively. The quantities corresponding to the incident, reflected and refracted waves will be denoted respectively by the subscripts "in", "refl" and "refr". The quantities corresponding to the medium in which the incident and reflected waves propagate will be assigned the subscript 1, while the subscript 2 will be used for the medium in which the refracted wave propagates. In particular, the permittivity of these two media is given by ε_1 and ε_2.

The electric field strengths of the incident, reflected and refracted plane waves are given by the formulas

$$\begin{aligned} \mathbf{E}_{\text{in}} &= \mathbf{E}_{\text{in}}^{(0)} e^{-i(\omega_{\text{in}} t - \mathbf{k}_{\text{in}} \cdot \mathbf{r})}, \\ \mathbf{E}_{\text{refl}} &= \mathbf{E}_{\text{refl}}^{(0)} e^{-i(\omega_{\text{refl}} t - \mathbf{k}_{\text{refl}} \cdot \mathbf{r})}, \\ \mathbf{E}_{\text{refr}} &= \mathbf{E}_{\text{refr}}^{(0)} e^{-i(\omega_{\text{refr}} t - \mathbf{k}_{\text{refr}} \cdot \mathbf{r})}. \end{aligned} \tag{16.2}$$

The wave numbers are connected with the velocities of the wave propagation in these media through the relations

$$k_{\text{in}} = \omega_{\text{in}}/v_1, \quad k_{\text{refl}} = \omega_{\text{refl}}/v_1, \quad k_{\text{refr}} = \omega_{\text{refr}}/v_2, \tag{16.3}$$

where $v_1 = (\varepsilon_1 \mu_1)^{-1}$, $v_2 = (\varepsilon_2 \mu_2)^{-1}$ represent the velocities of the wave propagation in the first and second media. The expressions for the magnetic induction vector are obtained from the general rule: the vector $\mathbf{B}$ is perpendicular to $\mathbf{E}$, oscillates in the same phase as $\mathbf{E}$ (in dielectrics), and has a magnitude $|\mathbf{B}| = |\mathbf{E}|/v$, where v is the velocity of wave propagation.

The laws of reflection and refraction are obtained as a consequence of the application of boundary conditions (16.1) to vectors (16.2) and to the corresponding magnetic field vectors.

CONSTANCY OF THE WAVE FREQUENCY UPON REFLECTION AND REFRACTION. The boundary conditions for the tangential components of the electric field strength have the form

$$\begin{aligned} &[\mathbf{E}_{\text{in}}^{(0)} e^{-i(\omega_{\text{in}} t - \mathbf{k}_{\text{in}} \cdot \mathbf{r})} + \mathbf{E}_{\text{refl}}^{(0)} e^{-i(\omega_{\text{refl}} t - \mathbf{k}_{\text{refl}} \cdot \mathbf{r})}]_\tau \\ &\quad = [\mathbf{E}_{\text{refr}}^{(0)} e^{-i(\omega_{\text{refr}} t - \mathbf{k}_{\text{refr}} \cdot \mathbf{r})}]_\tau, \end{aligned} \tag{16.4}$$

where the radius vector has an arbitrary origin. The radius vector can be presented in the form $\mathbf{r} = \mathbf{r}_n + \mathbf{r}_\tau$, where $\mathbf{r}_n$ and $\mathbf{r}_\tau$ are the normal (to the surface) and tangential components of the radius vector. In other words, $\mathbf{r}_n$ joins the origin of the radius vector along the normal with the corresponding point on the interface, while $\mathbf{r}_\tau$ lies on the interface and joins the tip of the vector $\mathbf{r}_n$ with the tip of the vector $\mathbf{r}$. The equality $\mathbf{k} \cdot \mathbf{r} = \mathbf{k}_n \cdot \mathbf{r}_n + \mathbf{k}_\tau \cdot \mathbf{r}_\tau$, where $\mathbf{k}_n$ and $\mathbf{k}_\tau$ are the normal and tangential components of the wave vector, means that as we go over from one point on the surface to another, the scalar product $\mathbf{k}_n \cdot \mathbf{r}_n$ remains constant and all the variation in the value of $\mathbf{k} \cdot \mathbf{r}$ is reduced to the variation in $\mathbf{k}_\tau \cdot \mathbf{r}_\tau$. Hence, for reasons which will become clear soon, it is expedient to choose the origin of the radius vectors at a point on the interface of the two media. For the rest, this point is arbitrary. To avoid complication of the formulas, we shall denote, as before, the radius vector $\mathbf{r}$ without the subscript "τ" although it lies on the interface. Equality (16.4) can be identically satisfied for

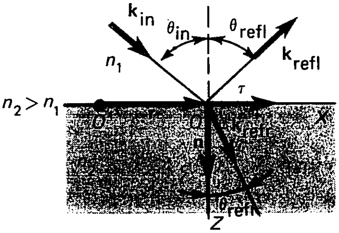


Fig. 58. Refraction of light for $n_2 > n_1$.

arbitrary and independent variations of t and r only if

$$\omega_{in} t = \omega_{refl} t = \omega_{refl} t, \tag{16.5}$$

$$\mathbf{k}_{in} \cdot \mathbf{r} = \mathbf{k}_{refl} \cdot \mathbf{r} = \mathbf{k}_{refr} \cdot \mathbf{r}. \tag{16.6}$$

It follows from (16.5) that

$$\underline{\overline{\omega_{in} = \omega_{refl} = \omega_{refr},}} \tag{16.7}$$

i. e. the frequency of an electromagnetic wave does not change as a result of reflection or refraction.

THE PLANE OF INCIDENT, REFLECTED AND REFRACTED RAYS. We shall call the direction characterizing the wave vector k a ray. We choose the origin of the vector r on the interface in such a way that this vector is perpendicular to the vector $\mathbf{k}_{in}$, i. e. the condition

$$\mathbf{k}_{in} \cdot \mathbf{r} = 0, \tag{16.8}$$

is satisfied. From (16.6), we obtain

$$\mathbf{k}_{refl} \cdot \mathbf{r} = \mathbf{k}_{refr} \cdot \mathbf{r} = 0, \tag{16.9}$$

which means that the vectors $\mathbf{k}_{refl}$ and $\mathbf{k}_{refr}$ are also perpendicular to r. This shows that the wave vectors of the incident, reflected and refracted rays all lie in the same plane.

RELATION BETWEEN THE ANGLES OF INCIDENCE, REFLECTION AND REFRACTION. Let us fix the origin of the rectangular Cartesian coordinate system at the point of incidence of a ray. We direct the Z-axis towards the medium in which the refracted ray propagates, and assume that the XZ-plane coincides with the plane in which the incident, reflected and the refracted rays lie (Fig. 58). The vectors $\mathbf{k}_{in}$, $\mathbf{k}_{refl}$ and $\mathbf{k}_{refr}$ are applied at the point O. (To simplify the drawing, these vectors have been plotted separately from this point along straight lines characterizing the propagation of these waves.) The unit vector n is directed into the second medium along the normal to the interface. The unit vector τ lies on the interface along the X-axis, and is a tangential vector. It should be recalled that the unit vectors are dimensionless. The angles θ_{in}, θ_{refl} and θ_{refr} are measured from perpendiculars to the interface and are respectively called the angles of incidence, reflection and refraction. The plane containing the incident wave vector and the normal to the interface at the point of incidence of the ray is called the plane of incidence of the ray.

The origin of the vector r is fixed at the point O' on the negative side of the X-axis of the coordinate system. Consequently, the vector $\mathbf{r} = r\tau$ is oriented along the positive direction of the X-axis. It can be seen from Fig. 60 that

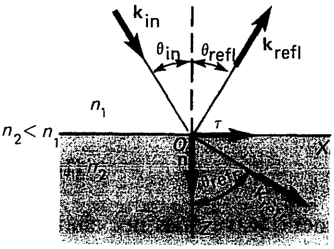


Fig. 59. Refraction of light for $n_2 < n_1$.

$$\begin{aligned} \mathbf{k}_{\text{in}} \cdot \mathbf{r} &= \mathbf{k}_{\text{in}} \cdot \tau r = k_{\text{in}} r \sin\theta_{\text{in}}, \\ \mathbf{k}_{\text{refl}} \cdot \mathbf{r} &= \mathbf{k}_{\text{refl}} \cdot \tau r = \mathbf{k}_{\text{refl}} r \sin\theta_{\text{refl}}, \\ \mathbf{k}_{\text{refr}} \cdot \mathbf{r} &= \mathbf{k}_{\text{refr}} \cdot \tau r = k_{\text{refr}} r \sin\theta_{\text{refr}}. \end{aligned} \tag{16.10}$$

This gives [see (16.6)]

$$k_{\text{in}} \sin\theta_{\text{in}} = k_{\text{refl}} \sin\theta_{\text{refl}} = k_{\text{refr}} \sin\theta_{\text{refr}}. \tag{16.11}$$

These relations can be simplified by taking into account (16.3):

$$\sin\theta_{\text{in}}/v_1 = \sin\theta_{\text{refl}}/v_1 = \sin\theta_{\text{refr}}/v_2. \tag{16.12}$$

It follows from (16.12) that $\sin\theta_{\text{in}} = \sin\theta_{\text{refl}}$, i.e.

$$\theta_{\text{in}} = \theta_{\text{refl}}, \tag{16.13}$$

$$\sin\theta_{\text{in}}/\sin\theta_{\text{refr}} = v_1/v_2 = n_{12}, \tag{16.14}$$

where $n_{12} = v_1/v_2$ is the refractive index of the second medium relative to the first. The refractive indices of the first and second media relative to vacuum are, by definition, given by

$$n_1 = c/v_1, \quad n_2 = c/v_2. \tag{16.15}$$

Hence

$$n_{12} = n_2/n_1. \tag{16.16}$$

Equation (16.13) shows that the angle of reflection is equal to the angle of incidence, while relation (16.14) expresses the Snell law of refraction: the ratio of the sine of the angle of incidence to the sine of the angle of refraction is equal to the refractive index of the medium containing the refracted ray relative to the medium containing the incident ray.

The refractive index depends on the wavelength and temperature and (for gases) on pressure also. For example, the refractive index of water at 20 °C is equal to 1.3308, 1.3330, 1.3374 and 1.3428 for wavelengths 678.0, 589.3, 480.0 and 404.7 nm respectively. If a very high accuracy is not required for the refractive index we can use the value $n = 1.33$. Under normal conditions, the refractive index of gases differs from unity by 10^{-3} or 10^{-4}. For example, at 0 °C and under atmospheric pressure, the refractive index n of nitrogen is 1.000297, and the corresponding values for oxygen and air are 1.000272 and 1.000292 respectively. Usually, the value 1.0003 is used for air. The refractive index of flint glass lies between 1.6 and 1.9, while for crown glass its value lies between 1.5 and 1.6. The refractive index of diamond is 2.4.

Figures 58 and 59 show the path of rays for $n_2 > n_1$ and $n_2 < n_1$. In the latter case, it can be seen that for a certain angle of incidence $\theta_{in} = \theta_{lim}$, called the limiting angle, the angle of refraction becomes equal to $\pi/2$, i. e. the refracted ray moves along the interface (Fig. 60), and there is no refracted ray in the second medium. Taking (16.14) into account, we obtain from (16.16)

$$\sin\theta_{lim} = n_2/n_1 \quad (n_2 < n_1). \tag{16.17}$$

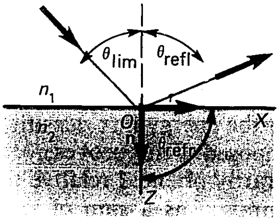


Fig. 60. The angle of refraction is equal to $\pi/2$ for limiting angle of incidence.

For angles of incidence exceeding θ_{lim}, Eq. (16.14) does not have a solution in the domain of real values of the angle θ_{refr}. Hence the situation arising in this case cannot be represented in the same way as in Figs. 58 and 59. This case requires a special analysis (see Sec. 17).

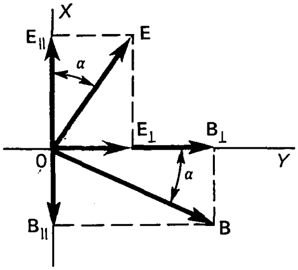


Fig. 61. Decomposition of a plane wave into two waves polarized in mutually perpendicular planes.

DECOMPOSITION OF A PLANE WAVE INTO TWO WAVES WITH MUTUALLY PERPENDICULAR LINEAR POLARIZATIONS. For a more clear presentation of the energy relations for reflection and refraction, it is expedient to split the general case of the incident wave, where the vector E_{in} is inclined at an arbitrary angle to the plane of incidence into two cases: when the vector E_{in} lies in the plane of incidence, and when it is perpendicular to this plane. For this purpose, we must prove that a plane wave can be presented as the sum of plane waves with mutually perpendicular polarizations, moreover, the sum of the energy flux densities of these two waves should be equal to the energy flux density of the initial wave. This relation does not follow directly from the principle of superposition.

We shall use the subscripts $\parallel$ and $\perp$ to denote the vector components lying in the plane of incidence and in a plane perpendicular to it. Since the reflected and refracted rays also lie in the plane of incidence, the electric field strength vectors for the incident, reflected and refracted waves have the following form in accordance with the principle of superposition:

$$\begin{gathered}\mathrm{E}_{in} = \mathrm{E}_{in\parallel} + \mathrm{E}_{in\perp}, \quad \mathrm{E}_{refl} = \mathrm{E}_{refl\parallel} + \mathrm{E}_{refl\perp}, \\ \mathrm{E}_{refr} = \mathrm{E}_{refr\parallel} + \mathrm{E}_{refr\perp}.\end{gathered} \tag{16.18}$$

The magnetic induction vectorsB can be also represented in a similar manner. Figure 61 shows this decomposition of field vectors. The *Z*-axis is perpendicular to the plane of the figure and is directed away from us. The wave propagates along the *Z*-axis. The *XZ*-plane is the plane of incidence. It can be seen from Fig. 61 that the following relations exist between the components of the field vectors:

$$\begin{gathered}E_{\parallel}/B_{\perp} = E\cos\alpha/(B\cos\alpha) = E/B, \\ E_{\perp}/B_{\parallel} = E\sin\alpha/(B\sin\alpha) = E/B.\end{gathered} \tag{16.19}$$

Moreover, it is obvious that the vectors ($E_\parallel$, $B_\perp$, k) and ($E_\perp$, $B_\parallel$, k) form right-handed triples of vectors. Hence we conclude that these vectors themselves form plane waves. The equality $E^2 = E_\parallel^2 + E_\perp^2$ shows that the energy flux density of the initial wave is equal to the sum of the energy flux densities of the waves into which it is decomposed. This proves that a plane wave whose vector E is arbitrarily oriented relative to the plane of incidence can be decomposed as a sum of two waves, one of which has its electric field strength vector in the plane of incidence, and the other vector in the plane perpendicular to it. By studying the behaviour of these waves upon reflection and refraction through a direct application of the principle of superposition and the additivity of the energy flux density, we obtain all the characteristics of the wave with an arbitrary orientation of the electric field strength vector.

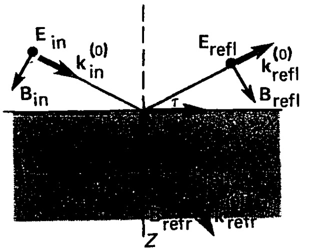


Fig. 62. Electric field strength of the incident wave is perpendicular to the plane of incidence.

S POLARIZACIJA →

VECTOR E IS PERPENDICULAR TO THE PLANE OF INCIDENCE (Fig. 62). The vectors $k_{in}^{(0)}$, $k_{refl}^{(0)}$ and $k_{refr}^{(0)}$ are unit vectors, i.e. the wave vectors are connected with these vectors through the relations $k_{in} = k_{in}^{(0)} k_{in}$, etc. The values of the remaining quantities are the same as in Fig. 60. We shall assume that E_{in} is directed towards us. To simplify the formulas, we shall omit the subscript ⊥, since we are considering strengths perpendicular to the plane of incidence. If the vector E_{in} is directed towards us, then the vector B_{in} will be directed as shown in Fig. 62. The direction of vectors E_{refl} and E_{refr} is not known beforehand. Hence we can solve the problem by assigning arbitrary directions to these vectors. If the solution gives the negative sign of these vectors, they will be opposite to the assigned direction. We shall assume that E_{refl} and E_{refr} are also directed towards us. The vectors B_{refl} and B_{refr} will then be oriented as shown in Fig. 62.

The boundary conditions (16.1b) for the continuity of the tangential components of the electric and magnetic field strengths have the form

$$E_{in} + E_{refl} = E_{refr}, \tag{16.20a}$$

$$(H_{in} + H_{refl})\tau = H_{refr}\tau. \tag{16.20b}$$

For subsequent transformations, it is convenient to represent the vector H in terms of the vector E:

$$H = k^0 \times E/Z, \tag{16.21}$$

where

$$Z = \sqrt{\mu/\varepsilon} \tag{16.22}$$

is the wave impedance of the medium. Note that for vacuum, $\sqrt{\mu_0/\varepsilon_0} = 377\ \Omega$. Considering that the vector E_{in} is perpendicular to the plane of the figure and is directed towards us (Fig. 62), we conclude that

$$\tau = \mathrm{E_{in}} \times \mathrm{n}/E_{\mathrm{in}}. \tag{16.23}$$

Hence the tangential component H of the incident, reflected and refracted waves can be presented in the form

$\vec{H}\cdot\vec{\tau} = -\frac{E}{Z}(\vec{k}^{0},\vec{n})$

$$\mathrm{H}\cdot\tau = (\mathrm{k}^{(0)} \times \mathrm{E})(\mathrm{E_{in}} \times \mathrm{n})/(ZE_{\mathrm{in}}) = -\,\mathrm{E}(\mathrm{k}^{(0)} \cdot \mathrm{n})/Z, \tag{16.24}$$

where H, E, Z and k stand for the corresponding quantities for the incident, reflected and refracted waves.

While carrying out the transformations in (16.24), we have used the familiar formula from vector algebra:

$$(\mathrm{A} \times \mathrm{B})(\mathrm{C} \times \mathrm{D}) = (\mathrm{A}\cdot \mathrm{C})(\mathrm{B}\cdot \mathrm{D}) - (\mathrm{A}\cdot \mathrm{D})(\mathrm{B}\cdot \mathrm{C}), \tag{16.25}$$

and considered that $\mathrm{k}^0 \cdot \mathrm{E_{in}} = 0$, $\mathrm{E}\cdot \mathrm{E_{in}} = EE_{\mathrm{in}}$ for all waves. The wave impedance of the medium for the incident and reflected waves is Z_1, and for the refracted wave it is equal to Z_2. Taking (16.24) into account, the boundary conditions (16.21) can be presented in the form

$$(1/Z_1)[E_{\mathrm{in}}(\mathrm{k}_{\mathrm{in}}^{(0)}\,\mathrm{n}) + E_{\mathrm{refl}}(\mathrm{k}_{\mathrm{refl}}^{(0)}\mathrm{n})] = (1/Z_2)E_{\mathrm{refr}}(\mathrm{k}_{\mathrm{refr}}^{(0)} \cdot \mathrm{n}). \tag{16.26}$$

* ознака (subscript) ⊥
означава да ради о
случају у коем је $\vec{E}$
нормално на упадну раван

FRESNEL'S FORMULAS FOR THE PERPENDICULAR COMPONENTS OF THE FIELD VECTOR. Equations (16.20a) and (16.26) can be written as a system of two equations in unknown ratios $E_{\mathrm{refl}}/E_{\mathrm{in}}$ and $E_{\mathrm{refr}}/E_{\mathrm{in}}$:

$$1 + E_{\mathrm{refl}}/E_{\mathrm{in}} = E_{\mathrm{refr}}/E_{\mathrm{in}}, \tag{16.27a}$$

$$(\mathrm{k}_{\mathrm{in}}^{(0)}\mathrm{n}) + (E_{\mathrm{refl}}/E_{\mathrm{in}})(\mathrm{k}_{\mathrm{refl}}^{(0)}\mathrm{n}) = (Z_1/Z_2)(\mathrm{k}_{\mathrm{refr}}^{(0)}\mathrm{n})(E_{\mathrm{refr}}/E_{\mathrm{in}}). \tag{16.27b}$$

The solution of this system is given by

*

$$(E_{\mathrm{refl}}/E_{\mathrm{in}})_{\perp} = [Z_2(\mathrm{k}_{\mathrm{in}}^{(0)}\cdot \mathrm{n}) - Z_1(\mathrm{k}_{\mathrm{refr}}^{(0)}\cdot \mathrm{n})]/[Z_2(\mathrm{k}_{\mathrm{in}}^{(0)}\cdot \mathrm{n}) + Z_1(\mathrm{k}_{\mathrm{refr}}^{(0)}\cdot \mathrm{n})], \tag{16.28a}$$

$$(E_{\mathrm{refr}}/E_{\mathrm{in}})_{\perp} = 2Z_2(\mathrm{k}_{\mathrm{in}}^{(0)}\cdot \mathrm{n})/[Z_2(\mathrm{k}_{\mathrm{in}}^{(0)}\cdot \mathrm{n}) + Z_1(\mathrm{k}_{\mathrm{refr}}^{(0)}\cdot \mathrm{n})], \tag{16.28b}$$

where $\mathrm{k}_{\mathrm{refl}}^{0}\cdot \mathrm{n} = -\,\mathrm{k}_{\mathrm{in}}^{0}\cdot \mathrm{n}$, since the angle of incidence is equal to the angle of reflection.

In accordance with (16.14), the angle of refraction is defined by the relation

$$\sin\theta_{\mathrm{refr}} = (n_1/n_2)\sin\theta_{\mathrm{in}}. \tag{16.29}$$

For $n_2 > n_1$, this equation in θ_{refr} has a real solution for all angles of incidence θ_{in}. If, however, $n_2 < n_1$, it will have a real solution only for angles of incidence that are smaller than the critical angle [see (16.17)]. Hence formulas (16.28) can be conveniently expressed in terms of the angles of incidence and refraction. In view of the fact that $\mathbf{k}_{\text{in}}^0 \cdot \mathbf{n} = \cos\theta_{\text{in}}$, $\mathbf{k}_{\text{refr}}^0 \cdot \mathbf{n} = \cos\theta_{\text{refr}}$, we can write (16.28) in the form

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_\perp = \frac{Z_2\cos\theta_{\text{in}} - Z_1\cos\theta_{\text{refr}}}{Z_2\cos\theta_{\text{in}} + Z_1\cos\theta_{\text{refr}}}$$

$$= \frac{\cos\theta_{\text{in}} - (\mu_1/\mu_2)\sqrt{n_{12}^2 - \sin^2\theta_{\text{in}}}}{\cos\theta_{\text{in}} + (\mu_1/\mu_2)\sqrt{n_{12}^2 - \sin^2\theta_{\text{in}}}}, \tag{16.30a}$$

$$\left(\frac{E_{\text{refr}}}{E_{\text{in}}}\right)_\perp = \frac{2Z_2\cos\theta_{\text{in}}}{Z_2\cos\theta_{\text{in}} + Z_1\cos\theta_{\text{refr}}}$$

$$= \frac{2\cos\theta_{\text{in}}}{\cos\theta_{\text{in}} + (\mu_1/\mu_2)\sqrt{n_{12}^2 - \sin^2\theta_{\text{in}}}}. \tag{16.30b}$$

Using the relation

$$Z_1/Z_2 = \mu_1\sin\theta_{\text{in}}/(\mu_2\sin\theta_{\text{refr}}), \tag{16.31}$$

we can write the above formulas in a different form:

$$(E_{\text{refl}}/E_{\text{in}})_\perp = (\mu_2\tan\theta_{\text{refr}} - \mu_1\tan\theta_{\text{in}})/(\mu_2\tan\theta_{\text{refr}}$$

$$+ \mu_1\tan\theta_{\text{in}}), \tag{16.32a}$$

$$(E_{\text{refr}}/E_{\text{in}})_\perp = 2\mu_2\tan\theta_{\text{refr}}/(\mu_2\tan\theta_{\text{refr}} + \mu_1\tan\theta_{\text{in}}). \tag{16.32b}$$

If the magnetic properties of the media on different sides of the interface are identical ($\mu_1 = \mu_2$), we obtain the following relations from (16.32a) and (16.32b):

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_\perp = -\frac{\sin(\theta_{\text{in}} - \theta_{\text{refr}})}{\sin(\theta_{\text{in}} + \theta_{\text{refr}})}, \tag{16.32c}$$

$$\left(\frac{E_{\text{refr}}}{E_{\text{in}}}\right)_\perp = \frac{2\cos\theta_{\text{in}}\sin\theta_{\text{refr}}}{\sin(\theta_{\text{in}} + \theta_{\text{refr}})} = \frac{(n_1/n_2)\sin 2\theta_{\text{in}}}{\sin(\theta_{\text{in}} + \theta_{\text{refr}})}. \tag{16.32d}$$

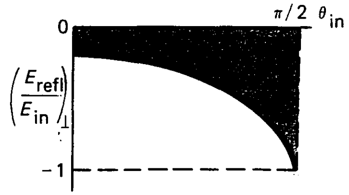


Fig. 63. Dependence of $(E_{refl}/E_{in})_\perp$ on the angle of incidence for $n_1 < n_2$.

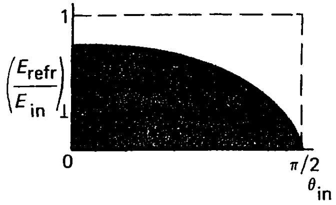


Fig. 64. Dependence of $(E_{refr}/E_{in})_\perp$ on the angle of incidence for $n_1 < n_2$.

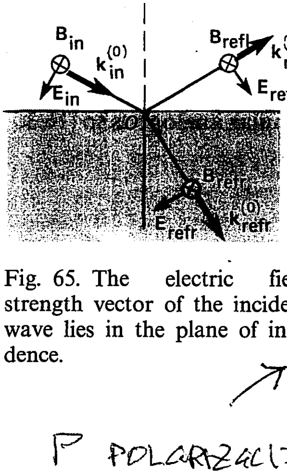


Fig. 65. The electric field strength vector of the incident wave lies in the plane of incidence.

↗

P POLARIZACIJA

These relations are called Fresnel's formulas for wave oscillations perpendicular to the plane of incidence. Fresnel (1788-1827) derived these relations in 1818 long before the emergence of Maxwell's electromagnetic theory of light, considering light as the oscillations of an elastic medium, the ether.

For normal incidence, Eqs. (16.32) lead to the relations ($\theta_{in} = 0$)

$$(E_{refl}/E_{in})_\perp = (\mu_2 n_1 - \mu_1 n_2)/(\mu_2 n_1 + \mu_1 n_2), \tag{16.33a}$$

$$(E_{refr}/E_{in})_\perp = 2\mu_2 n_1/(\mu_2 n_1 + \mu_1 n_2). \tag{16.33b}$$

For media with identical magnetic properties ($\mu_1 = \mu_2$) on two sides of the interface, we obtain

$$(E_{refl}/E_{in})_\perp = (n_1 - n_2)/(n_1 + n_2), \tag{16.33c}$$

$$(E_{refr}/E_{in})_\perp = 2n_1/(n_1 + n_2). \tag{16.33d}$$

Of course, formulas (16.33c) and (16.33d) can be also obtained from (16.32c) and (16.32d) by using the familiar mathematical methods for evaluating an indeterminate form in them.

Figures 63 and 64 show the variations in $(E_{refl}/E_{in})_\perp$ and $(E_{refr}/E_{in})_\perp$ with θ_{in} for $n_1 < n_2$ in accordance with formulas (16.32c) and (16.32d). It can be seen in Fig. 63 that $(E_{refl}/E_{in})_\perp$ is a negative quantity for all angles of incidence. This means that the direction of the vector $\mathbf{E}_{refl}$ is opposite to that shown in Fig. 62. In other words, when light is reflected at the interface with a medium having a higher refractive index, the phase of the reflected wave changes by π (the sign of the electric field strength is reversed). It can be seen from Fig. 64 that $(E_{refr}/E_{in})_\perp$ is always positive. Hence in this case the phase does not change as a result of refraction.

For incidence angles smaller than the limiting angle for $n_1 > n_2$, it can be seen that $\theta_{in} < \theta_{refr}$. Hence in formula (16.32c) $\sin(\theta_{in} - \theta_{refr}) < 0$ and $(E_{refl}/E_{in})_\perp > 0$. This means that when light is reflected at the interface with a medium having a smaller refractive index, the phase of vector **E** does not change. The refracted wave also has the same phase as the incident wave.

VECTOR **E** LIES IN THE PLANE OF INCIDENCE (Fig. 65). For all waves the vectors **B** are assumed to be directed away from us (behind the plane of the figure). In this case, we obtain instead of (16.21) and (16.23) the following relations:

$$\mathbf{E} = Z\mathbf{H} \times \mathbf{k}^{(0)}, \tag{16.34}$$

$$\boldsymbol{\tau} = \mathbf{n} \times \mathbf{H}_{in}/H_{in}. \tag{16.35}$$

It follows from here that

$$\mathbf{E}\cdot\boldsymbol{\tau} = -Z\mathbf{H}(\mathbf{k}^{(0)}\cdot\mathbf{n}) = -E(\mathbf{k}^{(0)}\cdot\mathbf{n}). \tag{16.36}$$

Hence the boundary conditions for the tangential components of the electric and magnetic fields, i. e.

$$(\mathrm{E}_{\mathrm{in}} + \mathrm{E}_{\mathrm{refl}}) \cdot \tau = \mathrm{E}_{\mathrm{refr}} \cdot \tau, \tag{16.37}$$

and

$$H_{\mathrm{in}} + H_{\mathrm{refl}} = H_{\mathrm{refr}}, \tag{16.38}$$

assume the form

$$E_{\mathrm{in}} (\mathrm{k}_{\mathrm{in}}^{(0)} \cdot \mathrm{n}) + E_{\mathrm{refl}} (\mathrm{k}_{\mathrm{refl}}^{(0)} \cdot \mathrm{n}) = (E_{\mathrm{in}} - E_{\mathrm{refl}}) (\mathrm{k}_{\mathrm{in}}^{(0)} \cdot \mathrm{n})$$

$$= E_{\mathrm{refr}} (\mathrm{k}_{\mathrm{refr}}^{(0)} \cdot \mathrm{n}), \tag{16.39}$$

$$(E_{\mathrm{in}} + E_{\mathrm{refl}})/Z_1 = E_{\mathrm{refr}}/Z_2. \tag{16.40}$$

FRESNEL'S FORMULAS FOR PARALLEL COMPONENTS OF FIELD VECTOR. Solving these equations in $E_{\mathrm{refl}}/E_{\mathrm{in}}$ and $E_{\mathrm{refr}}/E_{\mathrm{in}}$, we obtain in complete analogy with (16.28)

$$(E_{\mathrm{refl}}/E_{\mathrm{in}})_{\parallel} = [Z_1 (\mathrm{k}_{\mathrm{in}}^{(0)} \cdot \mathrm{n}) - Z_2 (\mathrm{k}_{\mathrm{refr}}^{(0)} \cdot \mathrm{n})]$$

$$/[Z_1 (\mathrm{k}_{\mathrm{in}}^{(0)} \cdot \mathrm{n}) + Z_2 (\mathrm{k}_{\mathrm{refr}}^{(0)} \cdot \mathrm{n})], \tag{16.41a}$$

$$(E_{\mathrm{refr}}/E_{\mathrm{in}})_{\parallel} = 2Z_2 (\mathrm{k}_{\mathrm{in}}^{(0)} \cdot \mathrm{n})/[Z_1 (\mathrm{k}_{\mathrm{in}}^{(0)} \cdot \mathrm{n}) + Z_2 (\mathrm{k}_{\mathrm{refr}}^{(0)} \cdot \mathrm{n})]. \tag{16.41b}$$

For real refraction angles (if $n_2 > n_1$) and for $\theta_{\mathrm{in}} < \theta_{\mathrm{refr}}$ (if $n_2 < n_1$), we obtain the following formulas analogous to (16.30)

$$\left(\frac{E_{\mathrm{refl}}}{E_{\mathrm{in}}}\right)_{\parallel} = \frac{Z_1 \cos\theta_{\mathrm{in}} - Z_2 \cos\theta_{\mathrm{refr}}}{Z_1 \cos\theta_{\mathrm{in}} + Z_2 \cos\theta_{\mathrm{refr}}}$$

$$= \frac{(\mu_1/\mu_2) n_{12}^2 \cos\theta_{\mathrm{in}} - \sqrt{n_{12}^2 - \sin^2\theta_{\mathrm{in}}}}{(\mu_1/\mu_2) n_{12}^2 \cos\theta_{\mathrm{in}} + \sqrt{n_{12}^2 - \sin^2\theta_{\mathrm{in}}}}, \tag{16.42a}$$

$$\left(\frac{E_{\mathrm{refr}}}{E_{\mathrm{in}}}\right)_{\parallel} = \frac{2Z_2 \cos\theta_{\mathrm{in}}}{Z_1 \cos\theta_{\mathrm{in}} + Z_2 \cos\theta_{\mathrm{refr}}}$$

$$= \frac{2n_{12} \cos\theta_{\mathrm{in}}}{(\mu_1/\mu_2) n_{12}^2 \cos\theta_{\mathrm{in}} + \sqrt{n_{12}^2 - \sin^2\theta_{\mathrm{in}}}}. \tag{16.42b}$$

For $\mu_1 = \mu_2$, we can transform these formulas with the help of (16.31):

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_{\parallel} = \frac{\sin 2\theta_{\text{in}} - \sin 2\theta_{\text{refr}}}{\sin 2\theta_{\text{in}} + \sin 2\theta_{\text{refr}}} = \frac{\tan(\theta_{\text{in}} - \theta_{\text{refr}})}{\tan(\theta_{\text{in}} + \theta_{\text{refr}})}, \tag{16.43a}$$

$$\left(\frac{E_{\text{refr}}}{E_{\text{in}}}\right)_{\parallel} = \frac{4\cos\theta_{\text{in}}\sin\theta_{\text{refr}}}{\sin 2\theta_{\text{in}} + \sin 2\theta_{\text{refr}}} = \frac{2\cos\theta_{\text{in}}\sin\theta_{\text{refr}}}{\sin(\theta_{\text{in}} + \theta_{\text{refr}})\cos(\theta_{\text{in}} - \theta_{\text{refr}})}. \tag{16.43b}$$

These formulas are called Fresnel's formulas for the electric field strength components of a wave lying in the plane of incidence. Together with (16.32), they give a complete solution of the problem of the behaviour of electromagnetic wave fields upon reflection and refraction at the interface of two dielectrics for an arbitrary orientation of the incident wave vectors relative to the plane of incidence.

For normal incidence ($\theta_{\text{in}} = 0$), these formulas are simplified:

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_{\parallel} = \frac{\mu_1 n_2 - \mu_2 n_1}{\mu_1 n_2 + \mu_2 n_1},$$

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_{\parallel} = \frac{n_2 - n_1}{n_2 + n_1}, \quad (\mu_1 = \mu_2), \tag{16.44a}$$

$$\left(\frac{E_{\text{refr}}}{E_{\text{in}}}\right)_{\parallel} = \frac{2\mu_2 n_1}{\mu_1 n_2 + \mu_2 n_1},$$

$$\left(\frac{E_{\text{refr}}}{E_{\text{in}}}\right)_{\parallel} = \frac{2n_1}{n_2 + n_1}, \quad (\mu_1 = \mu_2). \tag{16.44b}$$

At first glance, it seems that formula (16.44a) is in contradiction to formula (16.33a) since for normal incidence at the interface, the plane of incidence can be assumed to be oriented arbitrarily and hence there should be no difference in the formulas for parallel and perpendicular components, while (16.44a) and (16.33a) have opposite signs. However, there is no contradiction between these formulas in actual practice. As a matter of fact, formula (16.33a) was derived for the case when the vectors $\mathbf{E}_{\text{in}}$ and $\mathbf{E}_{\text{refl}}$ have the same direction in the case of normal incidence (see Fig. 64) and the ratio $E_{\text{in}}/E_{\text{refl}}$ is assumed to be positive. On the other hand, formulas (16.44a) were derived for the case in which the vectors $\mathbf{E}_{\text{in}}$ and $\mathbf{E}_{\text{refl}}$ were assumed to be directed opposite to each other with the normal incidence (Fig. 65), so that the ratio $E_{\text{refl}}/E_{\text{in}}$ was

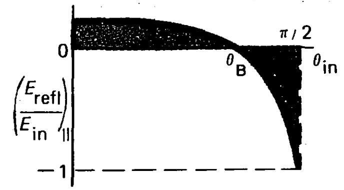


Fig. 66. Dependence of $(E_{refl}/E_{in})_{\|}$ on the angle of incidence for $n_1 < n_2$.

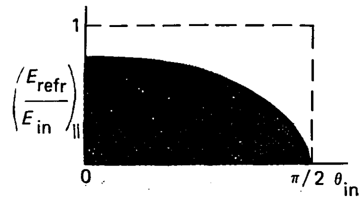


Fig. 67. Dependence of $(E_{refr}/E_{in})_{\|}$ on the angle of incidence for $n_1 < n_2$.

assumed to be negative. The difference in the signs of formulas (16.33a) and (16.44a) reflects this difference in the initial assumptions made for deriving these formulas.

Figures 66 and 67 show the plots of $(E_{refl}/E_{in})_{\|}$ and $(E_{refr}/E_{in})_{\|}$ as functions of the angle of incidence θ_{in} for $n_1 < n_2$ given by formulas (16.43). Figure 67 is similar to Fig. 64. Together, these figures show that for any orientations of the electric field strength for the incident wave, no phase differences arise as a result of refraction. Moreover, it can be seen that when the wave incidence is tangential to the interface ($\theta_{in} = \pi/2$), there is no refracted wave.

BREWSTER EFFECT. An analysis of the path of rays shown in Fig. 66 reveals that for waves whose electric field vector lies in the plane of incidence at an angle of incidence θ_B, there is no reflected wave at all. This phenomenon was discovered experimentally in 1815 by Brewster (1781-1868) and is called the Brewster effect. The angle θ_B is determined from formula (16.43a) under the condition $(E_{refl}/E_{in})_{\|} = 0$, i. e. when the denominator of the right-hand side of the equality turns to infinity. Thus, this angle is obtained from the condition $\theta_B + \theta_{refr\,B} = \pi/2$, where $\theta_{refr\,B}$ is the angle of refraction corresponding to $\theta_{in} = \theta_B$. According to the Snell law of refraction, $n_1 \sin\theta_B = n_2 \sin\theta_{refr\,B} = n_2 \sin(\pi/2 - \theta_B) = n_2 \cos\theta_B$. Hence we obtain the equality

$$\tan\theta_B = n_2/n_1 . \tag{16.45}$$

This formula is called the Brewster law and the angle at which there is no reflected wave is called the Brewster angle. It can be seen from Fig. 66 that the ratio $(E_{refl}/E_{in})_{\|}$ is positive for $\theta_{in} < \theta_B$ and negative for $\theta_{in} > \theta_B$. This means that while passing through the Brewster angle, the phase of the reflected wave abruptly changes by π.

Since for a wave incident at the Brewster angle the component of $\mathbf{E}_{in}$ lying in the plane of incidence is not reflected, the reflected wave contains only the electric field strength component perpendicular to the plane of incidence. In other words, the reflected light is completely polarized. Reflection at the Brewster angle is one of the ways of obtaining linearly polarized light.

The Brewster effect is due to the transverse nature of electromagnetic waves. Under the influence of the incident wave, the electrons of the medium begin to oscillate and emit secondary waves which combine with the primary wave. At the substitution length (see Sec. 15), the incident wave is completely replaced by a wave emitted by the oscillating electrons of the medium. The line of oscillation of electrons is collinear with the vector E of the wave. At the Brewster angle, when the angle

!

As a result of reflection, the phase of oscillations changes by π either for the electric field vector or for the magnetic field vector, in accordance with Fresnel's formulas. The phases of oscillations of wave vectors remain unchanged as a result of refraction.

In the case of incidence at the Brewster angle, only the wave whose vector E oscillates perpendicular to the plane of incidence is reflected.

between the reflected and the refracted waves would be 90°, the electrons of the medium which generate the refracted wave oscillate along a line parallel to the direction in which the reflected wave must propagate. Since an electron cannot emit an electromagnetic wave along the line of its oscillations, there is no reflected wave.

By studying the reflection of light at angles close to the Brewster angles, we can subject Fresnel's formulas to experimental verification with an extremely high degree of accuracy. This is due to the fact that the intensity of the wave which is reflected at the Brewster angle (and in which the vector oscillates in the plane of incidence) is equal to zero. Such a statement is convenient for experimental verification since there is no need to observe the angle of incidence precisely. It is sufficient to vary the angle of incidence continuously near the Brewster angle. As the angle of incidence becomes equal to the Brewster angle, the intensity of the wave with appropriate polarization must turn to zero. Moreover, measurements of zero type are usually more precise. According to the second statement, which is convenient for experimental verification, a wave reflected at the Brewster angle is exactly linearly polarized with its electric field vector oscillating perpendicular to the plane of incidence. In this case also, measurements of zero type are possible.

An analysis of the light reflected at the Brewster angle has shown that there are slight deviations from the predictions of Fresnel formulas. It has been found that there is no angle of incidence for which the reflected wave with its electric vector oscillating in the plane of incidence has zero intensity and for which the electric vector of the reflected wave oscillates along a line perpendicular to the plane of incidence. If the vector E of an incident linearly polarized wave is perpendicular to the plane of incidence, the wave reflected at the Brewster angle is found to be elliptically polarized, which is in contradiction with Fresnel's formulas predicting a linear polarization. It is also clear that the existence of a component of vector E of an elliptically polarized wave in the plane of incidence explains the absence of an angle at which the intensity of the reflected wave with appropriate polarization is zero.

The departure from Fresnel's formulas is explained by the fact that reflection and refraction take place not at the mathematical boundary of two media, but over a thin surface layer between them. The properties of the transition layer differ from the properties of the media which it demarcates. Hence the phenomena of reflection and refraction cannot be reduced to an abrupt and instantaneous change of parameters describing a wave, as was assumed in the derivation of

?

Write down the boundary conditions for the field vectors of a wave which completely define the laws of reflection and refraction. Which physical factors completely determine the value of the Brewster angle? Why is the incidence at the Brewster angle the most effective means of experimental verification of Fresnel's formulas?

Fresnel's formulas. The thickness of the transition layer responsible for reflection and refraction has the order of interatomic distance. In such a layer, the process of reflection and refraction cannot be described with the help of Maxwell's equations formulated under the assumption that the medium is continuous. We must use Maxwell's equations for vacuum and the transition layer must be represented by taking into account its molecular structure. Such an approach explains the departure from Fresnel's formulas. The deviations from Fresnel's formulas for angles differing from the Brewster angle are hard to observe.

RELATION BETWEEN THE PHASES OF REFLECTED AND REFRACTED WAVES. The direction of wave propagation is associated with the mutual orientation of vectors E and H in a wave through the right-hand screw rule. This means that upon reflection, the phase difference in the oscillations of E and H changes by π, while in the case of refraction the phase difference remains unchanged. The behaviour of the phases of waves upon reflection and refraction can be analyzed with the help of Fresnel's formulas and described as follows.

Under all conditions, refraction takes place without any change in the phase of the waves. In the case of reflection, the change in phase is determined by the conditions of reflection.

If a wave is incident on an interface from a medium with a lower optical density at any angles of incidence, the phase of the perpendicular component of the electric field strength in the reflected wave changes by π, i.e. $E_{\text{refl}\perp}$ and $E_{\text{in}\perp}$ have opposite phases at the interface. The phase of the parallel component $E_{\parallel}$ changes only for incidence angles smaller than the Brewster angle. For angles of incidence exceeding θ_B, the phase of the parallel component $E_{\text{refl}\parallel}$ remains unchanged (i.e. in this case the phase of the magnetic field strength changes by π).

When waves are incident at the interface from an optically denser medium at any angles of incidence, the phase of the perpendicular component of the electric field strength of the wave remains the same (i.e. in this case the phase of the magnetic field strength changes by π). The phase of the parallel component remains unchanged only for angles of incidence smaller than the Brewster angle. For angles of incidence exceeding θ_B, the phase of the parallel component of the electric field strength of the reflected wave changes by π (and hence the phase of the magnetic field vector in the reflected wave does not change).

DEGREES OF POLARIZATION. The variation in the polarization of light upon reflection or refraction is described by the degree of polarization:

$$P = (I_\perp - I_\parallel)/(I_\perp + I_\parallel). \tag{16.46}$$

For unpolarized light, $P = 0$. For completely polarized light, when the plane of oscillations of vector E is perpendicular to the plane of incidence, $P = 1$. If the vector E lies in the plane of incidence, then $P = -1$. With the help of definition (16.46), we can consider the polarization of light in the analysis of reflection and refraction. It must, however, be borne in mind that in the general theory of partially polarized light (see Sec. 30), the degree of polarization indicates the quantity (30.63) and not (16.46), although the two quantities are termed in the same way and have the same notation.

SEC. 17. TOTAL REFLECTION OF LIGHT

The total reflection of light is analyzed and the properties of a wave penetrating another medium are studied.

FORMULAS FOR ANGLES $\theta_{\text{in}} \geqslant \theta_{\text{lim}}$. Total reflection takes place for $n_1 > n_2$ and $\theta_{\text{in}} > \theta_{\text{lim}}$ [see Fig. 59 and (16.17)]. The most important physical characteristic of total reflection is the absence of a refracted wave. In this case, Eq. (16.14) does not have a solution in the domain of real values of the angle of refraction θ_{refr}. To analyze total reflection, we must express the laws of refraction through formulas that are also valid for $n_1 > n_2$ and $\theta_{\text{in}} > \theta_{\text{lim}}$. To begin with, we can represent Eqs. (16.11) in the form

$$k_{\text{in}} = k_{\text{refl}\,\tau} = k_{\text{refr}\,\tau}, \tag{17.1}$$

where $k_{\text{in}} = k_{\text{in}} \sin\theta_{\text{in}}$, $k_{\text{refl}\,\tau} = k_{\text{refl}} \sin\theta_{\text{refl}}$, $k_{\text{refr}\,\tau} = k_{\text{refr}} \times \sin\theta_{\text{refr}}$ are respectively the tangential components of the wave vectors of the incident, reflected and refracted waves. The relations between the normal components of the wave vectors of these waves (assuming that the positive normal is the one directed towards the second medium, see Fig. 59) are obtained from the obvious equalities

$$k_{\text{in}}^2 = k_{\text{in}\,\tau}^2 + k_{\text{in}\,n}^2,$$
$$k_{\text{refl}}^2 = k_{\text{refl}\,\tau}^2 + k_{\text{refl}\,n}^2,\quad k_{\text{refr}}^2 = k_{\text{refr}\,\tau}^2 + k_{\text{refr}\,n}^2 \tag{17.2}$$

in combination with relations (17.1):

$$k_{\text{refl}\,n} = -k_{\text{in}\,n},\quad k_{\text{refr}\,n}^2 = k_2^2 - k_{\text{in}\,\tau}^2, \tag{17.3}$$

where k_2 is the magnitude of the wave vector in the second medium (wave number), and the negative sign in the first of equalities (17.3) indicates that the reflected wave moves in a direction opposite to the positive direction of the normal, while the refracted wave moves in the same direction as the positive normal. We note that $k_{\text{in}\,\tau} = k_1 \sin\theta_{\text{in}}$, where k_1 is the wave number in the first medium. Under condition (17.1), Eqs. (17.3) assume the form

$$k_{\text{refr}\,\tau} = k_1 \sin\theta_{\text{in}}, \tag{17.4}$$

$$k^2_{\text{refr}\,n} = k_2^2 - k_1^2 \sin^2\theta_{\text{in}} = k_2^2[1 - (n_1/n_2)^2 \sin^2\theta_{\text{in}}]. \tag{17.5}$$

In this form, the formulas are valid for any values of the angle θ_{in} and of the ratio n_1/n_2. The angle θ_{refr} is determined from the relations

$$k_{\text{refr}\,\tau} = k_2 \sin\theta_{\text{refr}}, \quad k_{\text{refr}\,n} = k_2 \cos\theta_{\text{refr}}. \tag{17.6}$$

If this angle has a geometrical interpretation like the one shown in Figs. 58 and 59, its physical meaning is obvious: it is the angle of refraction. In the case of total reflection, this angle does not have a geometrical interpretation although formulas (17.4) and (17.5) remain valid and θ_{refr} is defined, as before, by equality (17.6).

Substituting $k_{\text{refr}\,\tau}$ and $k_{\text{refr}\,n}$ from (17.6) into formulas (17.4) and (17.5), we obtain the laws of refraction:

$$n_1 \sin\theta_{\text{in}} = n_2 \sin\theta_{\text{refr}}, \tag{17.7}$$

$$\cos^2\theta_{\text{refr}} = 1 - (n_1/n_2)^2 \sin^2\theta_{\text{in}}. \tag{17.8}$$

These two relations are in complete mutual agreement, since $\cos^2\theta_{\text{refr}} = 1 - \sin^2\theta_{\text{refr}}$. Formula (17.7) is identical with the Snell refraction law (16.14). Formula (17.8) can be used to find the value of $\cos\theta_{\text{refr}}$ not only when θ_{refr} has a geometrical meaning, but also when there is no such interpretation upon total reflection.

WAVE IN THE SECOND MEDIUM. We direct the coordinate axes as shown in Fig. 60. The electric field strengths in the incident and refracted waves are given by the formulas

$$\begin{aligned} \mathbf{E}_{\text{in}} &= \mathbf{E}_{\text{in}}^{(0)} \exp[-i(\omega t - \mathbf{k}_{\text{in}} \cdot \mathbf{r})] \\ &= \mathbf{E}_{\text{in}}^{(0)} \exp\{-i[\omega t - k_1(x \sin\theta_{\text{in}} + z \cos\theta_{\text{in}})]\}, \end{aligned} \tag{17.9}$$

$$\begin{aligned} \mathbf{E}_{\text{refr}} &= \mathbf{E}_{\text{refr}}^{(0)} \exp[-i(\omega t - \mathbf{k}_{\text{refr}} \cdot \mathbf{r})] \\ &= \mathbf{E}_{\text{refr}}^{(0)} \exp\{-i[\omega t - k_{\text{refr}\,x} x - k_{\text{refr}\,z} z]\}, \end{aligned} \tag{17.10}$$

where in accordance with (17.4) we can take $k_{\text{refr}\,x} = k_{\text{refr}\,\tau} = k_1 \sin\theta_{\text{in}}$. The value of $k_{\text{refr}\,\tau} = k_{\text{refr}\,n}$ is obtained from (17.5). Considering that the quantity in the square brackets in (17.5) is negative, we get

$$\begin{aligned} k^2_{\text{refr}\,n} = k^2_{\text{refr}\,z} &= -k_2^2[(n_1/n_2)^2 \sin\theta_{\text{in}} - 1] \\ &= -k_2^2[(\sin\theta_{\text{in}}/\sin\theta_{\text{lim}})^2 - 1] = -s^2, \end{aligned} \tag{17.11}$$

where s is a real quantity. Hence $k_{\mathrm{refr}\,z} = \pm is$, whose substitution into (17.10) leads to the appearance of the real multiplier $\exp(+sz)$. Consequently, the general solution $\mathrm{E}_{\mathrm{refr}}$ for the refracted wave consists of two parts, one of which contains the multiplier $\exp(sz)$ and the other contains $\exp(-sz)$. The multiplier $\exp(sz)$ increases to infinity as $z \to \infty$, which corresponds to an unlimited increase in the amplitude as we move in the second medium away from its boundary with the first medium. Obviously, this is unacceptable from physical considerations (the law of conservation of energy) and must be rejected. This leaves the term containing the multiplier $\exp(-sz)$, which has the form

$$\mathrm{E}_{\mathrm{refr}} = \mathrm{E}^{(0)}_{\mathrm{refr}}\, \mathrm{e}^{-sz} \exp[-i(\omega t - k_1 x \sin\theta_{\mathrm{in}})], \tag{17.12}$$

where

$$s = k_2\,[(n_1/n_2)^2 \sin^2\theta_{\mathrm{in}} - 1]^{1/2}$$
$$= (n_2\omega/c)\,[(\sin\theta_{\mathrm{in}}/\sin\theta_{\mathrm{lim}})^2 - 1]^{1/2}. \tag{17.13}$$

PENETRATION DEPTH. Formula (17.12) describes a wave propagating along the interface of two media in the positive direction of the X-axis. The amplitude of this wave decreases exponentially upon moving from the interface into the second medium. The amplitude decreases by a factor of e at a distance

$$\Delta = \frac{1}{s} = \frac{\lambda_2/(2\pi)}{[(n_1/n_2)^2 \sin^2\theta_{\mathrm{refr}} - 1]^{1/2}}, \tag{17.14}$$

called the penetration depth (see Sec. 19).

!

In the case of total reflection, a ray penetrates the second medium to a depth of the order of magnitude of the wavelength, and propagates in the second medium parallel to the interface with a phase velocity which is smaller than the phase velocity of the wave in the second medium. The ray returns to the first medium at a point which is displaced relative to the point of incidence.

PHASE VELOCITY. In order to determine the phase velocity of the refracted wave in the direction of the X-axis, we must assume, as usual, the phase of wave (17.12) to be constant and differentiate it with respect to time:

$$v_{2x} = \frac{\omega}{k_1 \sin\theta_{\mathrm{in}}} = \frac{c}{n_1 \sin\theta_{\mathrm{in}}} = \frac{(c/n_2)}{(n_1/n_2)\sin\theta_{\mathrm{in}}} = \left(\frac{\sin\theta_{\mathrm{lim}}}{\sin\theta_{\mathrm{in}}}\right) v_2, \tag{17.15}$$

where $v_2 = c/n_2$ is the phase velocity of a plane electromagnetic wave in the second medium. Since $\theta_{\mathrm{in}} > \theta_{\mathrm{lim}}$, the velocity v_{2x} is lower than v_2. The wave described by formula (17.12) is not a plane wave since its amplitude is not constant in a plane perpendicular to the direction of its propagation. The phase velocity of this wave is also not constant and depends on the properties of the medium as well as on the angle of incidence.

REFLECTED WAVE. The amplitude of the reflected wave can be determined with the help of formulas (16.28a) and (16.41a) which assume the following form when combined with relations (17.7), (17.8) and (17.13):

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_{\perp} = \frac{Z_2 \cos\theta_{\text{in}} - Z_1[1-(n_1/n_2)^2\sin^2\theta_{\text{in}}]^{1/2}}{Z_2 \cos\theta_{\text{in}} + Z_1[1-(n_1/n_2)^2\sin^2\theta_{\text{in}}]^{1/2}}$$

$$= \frac{Z_2 \cos\theta_{\text{in}} - iZ_1(s/k_2)}{Z_2 \cos\theta_{\text{in}} + iZ_1(s/k_2)}, \tag{17.16}$$

$$\left(\frac{E_{\text{refl}}}{E_{\text{in}}}\right)_{\parallel} = \frac{Z_1 \cos\theta_{\text{in}} - Z_2[1-(n_1/n_2)^2\sin^2\theta_{\text{in}}]^{1/2}}{Z_1 \cos\theta_{\text{in}} + Z_2[1-(n_1/n_2)^2\sin^2\theta_{\text{in}}]^{1/2}}$$

$$= \frac{Z_1 \cos\theta_{\text{in}} + iZ_2(s/k_2)}{Z_1 \cos\theta_{\text{in}} + iZ_2(s/k_2)}. \tag{17.17}$$

This gives

$$(E_{\text{refl}}/E_{\text{in}})_{\perp} = e^{2i\varphi_{\perp}}, \quad (E_{\text{refl}}/E_{\text{in}})_{\parallel} = e^{2i\varphi_{\parallel}} \tag{17.18}$$

where

$$\tan\varphi_{\perp} = -\frac{Z_1(s/k_2)}{Z_2\cos\theta_{\text{in}}}, \quad \tan\varphi_{\parallel} = -\frac{Z_2(s/k_2)}{Z_1\cos\theta_{\text{in}}}. \tag{17.19}$$

Consequently, the amplitudes of the electric field strength components of a plane wave parallel and perpendicular to the plane of incidence do not change in magnitude as a result of reflection, but undergo different phase changes. The equality of the magnitudes of amplitudes of the incident and reflected waves indicates the equality of their energy flux densities. Consequently, the energy flux of the incident wave is completely reflected and no part of the energy goes into the second medium. Hence such a reflection is called the total reflection. However, formula (17.12) shows that energy does penetrate the second medium. Considering that the energy is completely reflected, this penetration can be interpreted as follows: the wave and an appropriate part of energy penetrate the second medium through the interface to a very small depth Δ, move along the interface and then return to the first medium. The region where the energy enters the second medium is somewhat displaced with respect to the region where it returns to the first medium.

?

How can the penetration of a wave into the second medium be observed upon total reflection? Is total reflection possible in the case of a wave incident at the interface from an optically less dense medium? Does the phase velocity of a wave penetrating the second medium and propagating parallel to the interface depend on the angle of incidence?

The penetration of a wave into the second medium can be observed experimentally. The intensity of an electromagnetic wave in the second medium has a significant value only at distances smaller than the wavelength. Near the interface at which the incident light experiences total reflection on the side of the second medium, the luminescence of a thin fluorescent layer can be observed if the second medium contains some fluorescent material. This phenomenon is demonstrated in optics classes. The displacement of the exit point of the reflected light relative to the entrance point can also be demonstrated under laboratory conditions.

The penetration of an electromagnetic wave into the second medium is also observed in the microwave region, and in the region of longer waves. In the centimeter wave range, the aerial detector can be placed at different distances from the interface. This allows not only to verify that the wave enters the second medium, but also to analyze quantitatively the dependence of the wave amplitude on the distance to the interface.

Total reflection is responsible for the phenomenon of mirage when the ground is strongly heated. For example, if the road is very hot, the air temperature is the highest near the surface of the road and decreases away from it. Consequently, the refractive index of air has its minimum value near the surface and increases away from it. The rays incident at a very small angle at the road surface will then undergo total reflection. It may so happen that a driver concentrating his attention at an appropriate part of the road may see a distant vehicle in an inverted position. Similarly, total reflection of light from, say, clouds may create the impression that there are puddles on the surface of a hot road, and so on. Vertical temperature gradients in combination with convection currents may give rise to quite strange types of mirage.

SEC. 18. ENERGY RELATIONS FOR REFLECTION AND REFRACTION OF LIGHT

The method of transition from the characteristics of a wave vector upon reflection and refraction to the energy flux density characteristics is described.

ENERGY FLUX DENSITIES. The energy flux density of an electromagnetic wave, averaged over a period, is given by formula (15.10). In vector notation, this formula can be written in the form

$$\mathrm{S} = (1/2)v\varepsilon E^{(0)2}\mathrm{k}^{(0)}, \tag{18.1}$$

where $\mathrm{k}^{(0)}$ is a unit vector in the direction of the wave propagation, and the sign of averaging of S has been omitted for simplification of notation. Flux density (18.1) can be decomposed into normal and tangential components relative to the interface:

$$\mathbf{S} = \mathbf{S}_n + \mathbf{S}_\tau , \tag{18.2}$$

where

$$\mathbf{S}_n = (\mathbf{S} \cdot \mathbf{n})\, \mathbf{n}, \quad \mathbf{S}_\tau = (\mathbf{S} \cdot \boldsymbol{\tau})\, \boldsymbol{\tau} . \tag{18.3}$$

REFLECTION COEFFICIENT. The reflection coefficient is the ratio of the absolute magnitudes of the normal components of the energy flux density of the reflected and incident waves:

$$\rho = | \mathbf{S}_{\mathrm{refl}\, n} | / | \mathbf{S}_{\mathrm{in}\, n} | = | \mathbf{S}_{\mathrm{refl}} \cdot \mathbf{n} | / | \mathbf{S}_{\mathrm{in}} \cdot \mathbf{n} | . \tag{18.4}$$

Substituting (18.1) into (18.4) and considering that $| \mathbf{k}_{\mathrm{in}}^{(0)} \cdot \mathbf{n} | = | \mathbf{k}_{\mathrm{refl}}^{(0)} \cdot \mathbf{n} | = \cos\theta_{\mathrm{in}} = \cos\theta_{\mathrm{refl}}$, we obtain

$$\rho = | E_{\mathrm{refl}} / E_{\mathrm{in}} |^2 . \tag{18.5}$$

Considering the relations obtained earlier for the electric field strength in the reflected and incident waves for different orientations of $\mathbf{E}_{\mathrm{in}}$ relative to the plane of incidence, we can find the reflection coefficients $\rho_\perp$ and $\rho_\parallel$ for perpendicular and parallel orientations of the vector $\mathbf{E}_{\mathrm{in}}$ relative to the plane of incidence. Taking (16.30a) and (16.41a) into consideration, we can write on the basis of (18.5)

$$\rho_\perp = [(Z_2 \cos\theta_{\mathrm{in}} - Z_1 \cos\theta_{\mathrm{refr}})/(Z_2 \cos\theta_{\mathrm{in}} + Z_1 \cos\theta_{\mathrm{refr}})]^2 , \tag{18.6a}$$

$$\rho_\parallel = [(Z_1 \cos\theta_{\mathrm{in}} - Z_2 \cos\theta_{\mathrm{refr}})/(Z_1 \cos\theta_{\mathrm{in}} + Z_2 \cos\theta_{\mathrm{refr}})]^2 . \tag{18.6b}$$

The right-hand sides of these equalities can also be expressed through Fresnel's formulas (16.32a) and (16.43a). When the wave is normally incident on the interface, we obtain the following relation with the help of (16.33a) and (16.44a):

$$\rho_{\mathrm{norm}} = [(n_1 - n_2)/(n_1 + n_2)]^2 . \tag{18.7}$$

Let us estimate the reflection coefficient when light is normally incident from air to the surface of glass for which $n_2/n_1 \simeq 1.5$. Formula (18.7) shows that $\rho_{\mathrm{norm}} = 0.04$ in this case, which means that about 4% of the energy are reflected while the remaining 96% pass through glass.

TRANSMISSION COEFFICIENT. The energy flux density in a refracted wave is characterized by the transmission coefficient which is defined in the same way as (18.4):

$$\tau = | \mathbf{S}_{\mathrm{refr}} \cdot \mathbf{n} | / | \mathbf{S}_{\mathrm{in}} \cdot \mathbf{n} | . \tag{18.8}$$

Considering that

$k_{\text{in}\,n} = k_1 \cos\theta_{\text{in}}, \quad k_{\text{in}\,\tau} = k_1 \sin\theta_{\text{in}},$

$k_{\text{refr}\,n} = k_2 \cos\theta_{\text{refr}}, \quad k_{\text{refr}\,\tau} = k_2 \sin\theta_{\text{refr}},$

combining these relations and (18.1), we can transform (18.8) as follows:

$$\tau = \frac{Z_1 \cos\theta_{\text{refr}}}{Z_2 \cos\theta_{\text{in}}} \left| \frac{E_{\text{refr}}}{E_{\text{in}}} \right|^2 . \tag{18.9}$$

In the same way as for (18.6), we obtain the following relations from here by using formulas (16.30b) and (16.41b):

$$\tau_\perp = 4Z_1Z_2 \cos\theta_{\text{in}} \cos\theta_{\text{refr}}/(Z_2 \cos\theta_{\text{in}} + Z_1 \cos\theta_{\text{refr}})^2 , \tag{18.10a}$$

$$\tau_\parallel = 4Z_1Z_2 \cos\theta_{\text{in}} \cos\theta_{\text{refr}}/(Z_1 \cos\theta_{\text{in}} + Z_2 \cos\theta_{\text{refr}})^2 . \tag{18.10b}$$

For $\mu_1 = \mu_2$, the right-hand sides of these equations can be presented with the help of Fresnel's formulas as functions of angles of incidence and refraction only. For normal incidence, the transmission coefficient is equal to

$$\tau_{\text{norm}} = 4n_1n_2/(n_2 + n_1)^2 \quad (\mu_1 = \mu_2) . \tag{18.11}$$

LAW OF ENERGY CONSERVATION. A direct verification with the help of formulas (18.5), (18.6), (18.8) and (18.10) shows that the reflection and transmission coefficients satisfy the relation

$$\boxed{\rho + \tau = 1 ,} \tag{18.12}$$

which reflects the law of conservation of energy upon reflection and refraction.

POLARIZATION OF LIGHT UPON REFLECTION AND REFRACTION. Natural light is unpolarized. In view of a difference between $\rho_\perp$, $\tau_\perp$ and $\rho_\parallel$, $\tau_\parallel$, the reflected and refracted rays are partially polarized. Polarization upon reflection was experimentally observed in 1808 by Malus (1775-1812). Using a crystal of Iceland spar, he observed the birefringence of a sun ray (see Sec. 42) reflected from the surface of a glass plate. As the glass plate was rotated around the ray as the axis, Malus observed that the relative intensity of the two rays obtained as a result of birefringence changes. This shows that the sun ray is partially polarized as a result of reflection from the glass plate. Malus did not provide any theoretical explanation for polarization as a result of reflection. The polarization of light as a result of refraction was experimentally discovered in 1811 by Malus and Biot (1774-1862).

SEC. 19. PROPAGATION OF LIGHT IN CONDUCTING MEDIA

The method for taking into account the conductivity of a medium is described. The influence of conductivity on the wave properties is discussed.

COMPLEX PERMITTIVITY. The conductivity γ of a homogeneous conducting medium is nonzero, and hence Maxwell's equations (2.47)-(2.49) remain unchanged, while Eq. (2.46) can be replaced by

$$\nabla \times \mathrm{B} = \mu\gamma \mathrm{E} + \mu\varepsilon \partial \mathrm{E}/\partial t, \tag{19.1}$$

where we have used Ohm's law in differential form ($\mathrm{j} = \gamma \mathrm{E}$).

Substituting the expressions (2.50) and (2.51) into Eqs. (19.1), and (2.47)-(2.49), we obtain the following relation instead of (2.53):

$$-\mathrm{k} \times \mathrm{B} = \omega\mu[\varepsilon + i\gamma/\omega]\,\mathrm{E}, \tag{19.2}$$

while Eqs. (2.54)-(2.56) remain unchanged as a result of such a substitution. For $\gamma = 0$, these equations are transformed into equations for the propagation of light in a dielectric (see Sec. 15). It can be seen from (19.2) that all formulas for light propagation in a dielectric are transformed into formulas for light propagation in conducting media if the permittivity in these formulas is replaced by complex permittivity:

$$\varepsilon_\omega = \varepsilon + i\gamma/\omega. \tag{19.3}$$

In a dielectric medium, the wave number k, frequency ω and the permittivity ε are connected through relation (15.4) which can be written in the form

$$k^2 = \omega^2 \varepsilon\mu, \tag{19.4}$$

where $\varepsilon\mu = 1/v^2$. For a conducting medium, ε and k are replaced by ε_ω and k_ω which are determined in accordance with (19.4) from the equality

$$k_\omega^2 = \omega^2 \varepsilon_\omega \mu = \omega^2 \varepsilon\mu + i\omega\gamma\mu. \tag{19.5}$$

Representing k_ω as a complex number,

$$k_\omega = k + is, \tag{19.6}$$

we can write (19.5) in the form

$$k^2 + 2iks - s^2 = \omega^2 \varepsilon\mu + i\omega\gamma\mu. \tag{19.7}$$

Equating the real and imaginary parts on the left- and right-hand sides, we obtain

$$k^2 - s^2 = \omega^2 \varepsilon\mu \tag{19.8}$$

and

$$2ks = \omega\gamma\mu. \tag{19.9}$$

The solution of this system of equations has the form

$$k^2 = (\omega^2\varepsilon\mu/2)(\sqrt{1+(\gamma/\varepsilon\omega)^2}+1), \tag{19.10}$$

$$s^2 = (\omega^2\varepsilon\mu/2)(\sqrt{1+(\gamma/\varepsilon\omega)^2}-1). \tag{19.11}$$

PENETRATION DEPTH. When a plane wave propagates in a conducting medium, its field vectors are given by formulas (2.50) and (2.51) in which k has been replaced by k_ω; thus, for propagation in the positive direction of the Z-axis, these vectors have the form

$$\mathrm{E} = \mathrm{E}_0 \mathrm{e}^{-i(\omega t - k_\omega z)} = \mathrm{E}_0 \mathrm{e}^{-sz} \mathrm{e}^{-i(\omega t - kz)}, \tag{19.12}$$

$$\mathrm{B} = \mathrm{B}_0 \mathrm{e}^{-i(\omega t - k_\omega z)} = \mathrm{B}_0 \mathrm{e}^{-sz} \mathrm{e}^{-i(\omega t - kz)}, \tag{19.13}$$

where expression (19.6) has been used for k_ω. Thus, the amplitude of a plane wave decreases during its propagation, i.e. the wave is absorbed as it propagates. According to the general principle, the penetration depth is given by

$$\Delta = 1/s, \tag{19.14}$$

At large depths, the wave is practically absent.

For visible range and good conductors, formula (9.11) is considerably simplified. The conductivity γ of metals is $\sim 10^7$ S/m, and ε can be taken equal to ε_0. Consequently, in this case, we obtain

$$\gamma/(\varepsilon\omega) \simeq 2 \times 10^2 \gg 1. \tag{19.15}$$

Taking this relation into account, we can neglect unity in formula (19.11) in comparison with $\gamma/(\varepsilon\omega)$, and write the expression for s in the form

$$s = \sqrt{\omega\gamma\mu/2}. \tag{19.16}$$

Consequently, the penetration depth is

$$\boxed{\Delta = \sqrt{2/(\omega\gamma\mu)}.} \tag{19.17}$$

Since $\omega = 2\pi/(\lambda\sqrt{\varepsilon\mu})$, formula (19.17) can be also presented in the form

$$\Delta = \sqrt{\frac{\lambda}{\pi\gamma}} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/4} \left(\frac{\varepsilon_r}{\mu_r}\right)^{1/4}, \tag{19.18}$$

where $(\mu_0/\varepsilon_0)^{1/2} = 377\ \Omega$.

As an example, let us consider copper for which $\gamma = 5 \times 10^7$ S/m. For wavelengths in the visible range ($\lambda \simeq 1\ \mu$m), the penetration depth Δ is approximately equal to 4 nm, i.e. just a few thousandths of the wavelength. In this case, there is no sense in speaking of light propagation in the conductor. However, if the conductivity of the medium is not very high, light can be transmitted over a sufficiently large distance. The amplitude of the waves gradually decreases in this case.

PHYSICAL REASON BEHIND ABSORPTION. The physical reason behind the damping of electromagnetic waves in a conducting medium is the transformation of the electromagnetic energy of waves into Joule heat. The electric field strength of the wave induces conduction currents in the conducting medium, which heat the material according to Joule's law.

PHASE VELOCITY AND WAVELENGTH. Taking (19.10) into account, we obtain the following expression from (19.12) and (19.13) for the phase velocity:

$$v = \frac{\omega}{k} = \frac{1}{\sqrt{\mu\varepsilon}} \left\{ \frac{2}{[1 + (\gamma/\omega\varepsilon)^2]^{1/2} + 1} \right\}^{1/2}. \tag{19.19}$$

This velocity is lower than the velocity of light in a nonconductor with the same value of ε (for identical magnetic permeabilities). Thus, the conduction in a medium is responsible for a decreased phase velocity. The wavelength in a conducting medium is given by

$$\lambda = \frac{2\pi}{k} = \frac{2\pi}{\omega\sqrt{\mu\varepsilon}} \left\{ \frac{2}{[1 + (\gamma/\omega\varepsilon)^2]^{1/2} + 1} \right\}^{1/2}. \tag{19.20}$$

This value is also smaller than the wavelength in a nonconducting medium with the same values of ε and magnetic permeability.

It can be seen from (19.19) that the velocity of a wave in a conducting medium depends on frequency, i. e. the dispersion is observed during the propagation of the wave. Hence nonmonochromatic waves propagate in a conducting medium in the same way as through a dispersive medium.

RELATION BETWEEN THE PHASES OF FIELD VECTOR OSCILLATIONS. It is expedient to present the complex quantity k_ω in the exponential form:

$$k_\omega = |k_\omega| e^{i\varphi}, \quad |k_\omega| = \sqrt{k^2 + s^2}, \tag{19.21}$$

where $\tan\varphi = s/k$ [see (19.6)]. Formula (2.54) can be written in the form

$$\mathbf{B} = \frac{|k_\omega|}{\omega} e^{i\varphi} \mathbf{k}^{(0)} \times \mathbf{E}, \tag{19.22}$$

where $\mathbf{k}^{(0)}$ is the unit vector in the direction of wave propagation, in this case in the direction of the Z-axis. Vectors $\mathbf{E}$ and $\mathbf{B}$ are perpendicular to this axis.

The electric field strength of a wave is given by formula (19.12) in which $\mathbf{E}_0$ can be treated as a real quantity without any loss of generality, since the choice of the initial instant of time is always arbitrary. Substituting (19.12) into (19.22), we arrive at the equality

$$\mathbf{B} = \frac{|k_\omega|}{\omega} \mathbf{k}^{(0)} \times \mathbf{E}_0 \mathrm{e}^{-sz} \mathrm{e}^{-i(\omega t - kz - \varphi)}. \tag{19.23}$$

Taking the real parts of (19.12) and (19.23) for $\mathbf{E}$ and $\mathbf{B}$ respectively, we obtain the following relations (the real parts of $\mathbf{E}$ and $\mathbf{B}$ are denoted by the same letters):

$$\mathbf{E} = \mathbf{E}_0 \mathrm{e}^{-sz} \cos(\omega t - kz),$$

$$\mathbf{B} = \frac{|k_\omega|}{\omega} \mathbf{k}^{(0)} \times \mathbf{E}_0 \mathrm{e}^{-sz} \cos(\omega t - kz - \varphi), \tag{19.24}$$

which show that **the electric and magnetic vectors of the wave oscillate with a phase difference φ.** It follows from (19.21) that

$$\tan\varphi = s/k = \sqrt{1 + (\varepsilon\omega/\gamma)^2} - \sqrt{\varepsilon\omega/\gamma}, \tag{19.25}$$

which means that the angle φ is positive. Thus, **the phase of oscillations of B lags behind the phase of vector E.**

RELATION BETWEEN AMPLITUDES OF FIELD VECTORS. It follows from (19.12) and (19.23) that

$$|\mathbf{B}|/|\mathbf{E}| = |k_\omega|/\omega = \sqrt{\mu\varepsilon}\,[1 + (\gamma/(\varepsilon\omega))^2]^{1/4}. \tag{19.26}$$

Comparing this relation with (15.5) ($v = 1/\sqrt{\varepsilon\mu}$), we find that **in a conducting medium, the ratio of B to E is larger than in a nonconducting medium with the same permittivity and permeability.**

MEDIA WITH LOW CONDUCTIVITY. In a "bad" conductor $\gamma \to 0$ and hence $\omega\varepsilon/\gamma \to \infty$. All formulas containing $\gamma/(\omega\varepsilon) \ll 1$ can be approximated by expanding into a series in $\gamma/(\omega\varepsilon)$ and confining to the first nonvanishing term in the expansion:

$$k = \omega\sqrt{\mu\varepsilon}\left[1 + \frac{1}{8}\left(\frac{\gamma}{\omega\varepsilon}\right)^2\right], \quad S = \frac{\gamma}{2}\sqrt{\frac{\mu}{\varepsilon}}, \tag{19.27}$$

$$\tan\alpha = -\frac{1}{2}\frac{\gamma}{\omega\varepsilon},$$

$$|k_\omega| = \omega\sqrt{\mu\varepsilon}\left[1 + \frac{1}{4}\left(\frac{\gamma}{\omega\varepsilon}\right)^2\right], \tag{19.28}$$

$$v = \frac{1}{\sqrt{\mu\varepsilon}}\left[1 - \frac{1}{8}\left(\frac{\gamma}{\omega\varepsilon}\right)^2\right],$$

$$\lambda = \frac{2\pi}{\omega\sqrt{\mu\varepsilon}}\left[1 - \frac{1}{8}\left(\frac{\gamma}{\omega\varepsilon}\right)^2\right], \tag{19.29}$$

$$\frac{|B_0|}{|E_0|} = \sqrt{\mu\varepsilon}\left[1 + \frac{1}{4}\left(\frac{\gamma}{\omega\varepsilon}\right)^2\right]. \tag{19.30}$$

For $\gamma = 0$, these formulas are transformed into the corresponding expressions for a dielectric and quite clearly describe the effect of conductivity of the medium on the wave properties. It can be seen that the conductivity of a medium is responsible for a larger wave number, appearance of a phase shift between the field vectors, wave attenuation, decrease in the phase velocity and wavelength, and an increase in $|B|$ relative to $|E|$.

!

The electric conductivity of a medium is responsible for the absorption of wave energy in the medium. The conductivity of a medium is taken into account by the imaginary part of the complex permittivity.
The wavelength in a conducting medium is smaller than the wavelength in a nonconducting medium with the same values of permittivity and permeability.
The electric field vector and the magnetic induction vector of a wave oscillate in a conducting medium with a phase difference.

MEDIA WITH HIGH CONDUCTIVITY. In "good" conductors, $\gamma \to \infty$ and hence $\omega\varepsilon/\gamma \to 0$. As a result of expansion of the expressions into a series, we obtain

$$\begin{aligned} k &= \sqrt{\frac{\omega\gamma\mu}{2}}\left(1 + \frac{1}{2}\frac{\omega\varepsilon}{\gamma}\right), \\ s &= \sqrt{\frac{\omega\gamma\mu}{2}}\left(1 - \frac{1}{2}\frac{\omega\varepsilon}{\gamma}\right), \end{aligned} \tag{19.31}$$

$$\tan\varphi = \frac{\omega\varepsilon}{\gamma} - 1, \quad |k_\omega| = \sqrt{\omega\gamma\mu}, \tag{19.32}$$

$$v = \sqrt{\frac{2\omega}{\gamma\mu}}\left(1 - \frac{1}{2}\frac{\omega\varepsilon}{\gamma}\right), \quad \lambda = 2\pi\sqrt{\frac{2}{\omega\gamma\mu}}\left(1 - \frac{\omega\varepsilon}{\gamma}\right), \tag{19.33}$$

$$\frac{|B_0|}{|E_0|} = \sqrt{\frac{\gamma\mu}{\omega}} = \frac{\sqrt{\mu\varepsilon}}{\sqrt{\omega\varepsilon/\gamma}}. \tag{19.34}$$

It can be seen that the phase difference between field vectors in a good conductor is close to $-\pi/4$, while the phase velocity of a wave in a good conductor is much lower than in a dielectric. In a good conductor, the wave attenuation is very high.

Example 19.1. A wave with a circular frequency $\omega = 2.5 \times 10^{15}\ \mathrm{s}^{-1}$ propagates in a conducting medium. The intensity of the wave in the medium is found to decrease by

?

What is the physical reason behind the emergence of a phase difference in the oscillations of the electric field vector and the magnetic induction vector of a wave in a conducting medium?
Which physical factors are responsible for a decrease in the wavelength in a conducting medium in comparison to the wavelength in a nonconducting medium?

a factor of e over a path length $l = 20$ cm, and the wavelength λ in the medium is equal to 0.5 μm. Find the refractive index of the medium and its electric conductivity.

The refractive index is given by the formula

$$n = c/v = \lambda_0/\lambda = 2\pi c/(\omega\lambda) = 1.5.$$

Taking (19.12) into account, we obtain

$$2sl = 1\,, \tag{19.35}$$

where s is defined by Eq. (19.11). Dividing both sides of (19.11) by $(\omega\sqrt{\varepsilon_0\mu_0})^2 = (\omega/c)^2$, we obtain

$$[c/(2/\omega)]^2 = (\varepsilon_r/2)[\sqrt{1 + [\gamma/(\varepsilon\omega)]^2} - 1]\,, \tag{19.36}$$

where $\varepsilon_r = \varepsilon/\varepsilon_0$ is the relative permittivity of the medium, connected with the refractive index through the relation $n = \sqrt{\varepsilon_r}$ [the medium is assumed to be nonmagnetic $(\mu = \mu_0)$]. Considering that $c/(2/\omega) = 3 \times 10^8/(2 \times 0.2 \times 10^{15} \times 2.5) = 3 \times 10^{-7} \ll 1$, we obtain the inequality $[\gamma/(\varepsilon\omega)]^2 \ll 1$ from (19.36). On the basis of this inequality, the right-hand side of (19.36) can be expanded into a series in $[\gamma/(\varepsilon\omega)]^2$. Confining ourselves to the first nonvanishing term in this expansion, we obtain

$$\frac{c}{2/\omega} = \frac{\sqrt{\varepsilon_r}}{2}\,\frac{\gamma}{\varepsilon\omega} = \frac{1}{2}\,\frac{\gamma}{\varepsilon_0\omega}\,\frac{1}{\sqrt{\varepsilon_r}} = \frac{1}{2}\,\frac{\gamma}{\varepsilon_0\omega}\,\frac{1}{n}.$$

Consequently,

$$\gamma = \varepsilon_0 cn/l = 8.85 \times 10^{-12} \times 3 \times 10^8 \times 1.5/0.2 \text{ S/m} = 0.02 \text{ S/m}. \tag{19.37}$$

SEC. 20. REFLECTION OF LIGHT FROM A CONDUCTING SURFACE

The method for taking the conductivity of the medium into account during reflection is described.

BOUNDARY CONDITIONS. A wave incident on the surface of a conductor partly penetrates the conductor, where it is absorbed, and is partly reflected. For the sake of simplicity of formulas, we shall confine ourselves to normal incidence only.

We denote by ε_1 and μ_1 the permittivity and permeability of the medium from which a light wave is incident on the surface of a conductor. The corresponding quantities for the conductor are denoted by ε_2 and μ_2 (both the dielectric medium and the conductor are assumed to be nonmagnetic). The electric conductivity of the conductor is γ. The orientation of the coordinate system and the vectors characterizing the wave

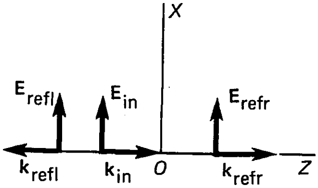


Fig. 68. Orientation of field vectors in the analysis of reflection of light at the surface of a conductor.

are shown in Fig. 68. The electric field strength of the waves can be written in the form

$$\mathbf{E}_{\text{in}} = \mathbf{i}_x E_{\text{in}}^{(0)} \mathrm{e}^{-i(\omega t - k_1 z)},$$
$$\mathbf{E}_{\text{refl}} = \mathbf{i}_x E_{\text{refl}}^{(0)} \mathrm{e}^{-i(\omega t + k_1 z)},$$
$$\mathbf{E}_{\text{refr}} = \mathbf{i}_x E_{\text{refr}}^{(0)} \mathrm{e}^{-i(\omega t - k_{2\omega} z)}, \tag{20.1}$$

where $\mathbf{i}_x$ is the unit vector in the positive direction of the X-axis and the quantities

$$k_1 = n_1 \omega / c, \quad k_{2\omega} = k_2 + i s_2 \tag{20.2}$$

are defined by relations (15.4) ($v = c/n_1$), (19.6), (19.10) and (19.11). The inductions of the corresponding magnetic fields are obtained from formula (16.22) in the form

$$\mathbf{B} = \mathbf{k} \times \mathbf{E}/\omega. \tag{20.3}$$

Taking into account the orientation of vectors (Fig. 68), we obtain from (20.3)

$$\mathbf{B}_{\text{in}} = \mathbf{i}_y k E_{\text{in}}/\omega, \quad \mathbf{B}_{\text{refl}} = -\mathbf{i}_y k_1 E_{\text{refl}}/\omega,$$
$$\mathbf{B}_{\text{refr}} = \mathbf{i}_y k_{2\omega} E_{\text{refr}}/\omega, \tag{20.4}$$

where $\mathbf{i}_y$ is the unit vector in the positive direction of the Y-axis.

The continuity of the tangential components of electric and magnetic field strengths is expressed through two equations:

$$E_{\text{in}}^{(0)} + E_{\text{refl}}^{(0)} = E_{\text{refr}}^{(0)},$$
$$k_1 (E_{\text{in}}^{(0)} - E_{\text{in}}^{(0)})/\mu_1 = k_{2\omega} E_{\text{refr}}^{(0)}/\mu_2. \tag{20.5}$$

!

As a result of reflection from a conducting surface, a phase difference emerges between the parallel and perpendicular components of the field vectors of a wave.

RELATION BETWEEN WAVE AMPLITUDES. Solving Eqs. (20.5) in unknown ratios $E_{\text{refl}}^{(0)}/E_{\text{in}}^{(0)}$ and $E_{\text{refr}}^{(0)}/E_{\text{in}}^{(0)}$, we obtain

$$\frac{E_{\text{refl}}^{(0)}}{E_{\text{in}}^{(0)}} = \frac{1 - (k_{2\omega}/k_1)(\mu_1/\mu_2)}{1 + (k_{2\omega}/k_1)(\mu_1/\mu_2)}, \tag{20.6}$$

$$\frac{E_{\text{refr}}^{(0)}}{E_{\text{in}}^{(0)}} = \frac{2}{1 + (k_{2\omega}/k_1)(\mu_1/\mu_2)}. \tag{20.7}$$

?

Which factors determine the polarization of a wave reflected from a conducting surface? What is the type of this polarization?
What is the dependence of the reflection coefficient on the electric conductivity and what is its value for good conductors?

Since $k_{2\omega}$ is a complex quantity, it follows from (20.6) and (20.7) that the phases of all the waves are shifted relative to one another. The phase shift can be determined from (20.6) and (20.7) by expressing their right-hand sides as the product of the modulus of a quantity and the phase factor in the same way as in (19.21).

REFLECTION COEFFICIENT. In accordance with definition (18.5), we obtain from (20.6) the following expression for the reflection coefficient:

$$\rho = \left|\frac{E_{\mathrm{refl}}^{(0)}}{E_{\mathrm{in}}^{(0)}}\right| = \frac{[1-\mu_1 k_2/(\mu_2 k_1)]^2 + [\mu_1 s_2/(\mu_2 k_1)]^2}{[1+\mu_1 k_2/(\mu_2 k_1)]^2 + [\mu_1 s_2/(\mu_2 k_1)]^2}, \tag{20.8}$$

where k_2 and s_2 are defined by formulas (19.10) and (19.11). Dividing the numerator and denominator of the above equation by $\mu_1 s_2/(\mu_2 k_1)$, we obtain the following equation:

$$\rho = \frac{\{[\mu_2 k_1/(\mu_1 s_2)] - (k_2/s_2)\}^2 + 1}{\{[\mu_2 k_1/(\mu_1 s_2)] + (k_2/s_2)\}^2 + 1}. \tag{20.9}$$

It can be seen from (19.10) and (19.11) that for an ideal conductor $\gamma \to \infty$, $s_2 \to \infty$, $k_2/s_2 \to 1$, $k_1/s_2 \to 0$. Hence we obtain from (20.9)

$$\rho \to 1. \tag{20.10}$$

Hence the reflection coefficient for very good conductors is close to unity.

CONNECTION BETWEEN REFLECTING AND ABSORBING PROPERTIES. It follows from formula (19.17) that the higher the conductivity, i.e. the better the conductor, the stronger the absorption of light by the conductor. It can be seen from (20.10) that the higher the conductivity, the better the reflecting properties of the conductor. Hence it can be stated that the better light is reflected by a conductor, the better it is absorbed by the same conductor. This leads to a difference in the colour of a conductor as viewed in reflected and transmitted lights (see Sec. 15).

PROBLEMS

3.1. Find the expression for the energy density per unit volume of a plane wave propagating in a conductor in the direction of the Z-axis.

3.2. Find the expression for the refractive index of a very good conductor.

3.3. Find the ratio $\langle w_{\mathrm{m}}\rangle/\langle w_{\mathrm{e}}\rangle$ of the energy density per unit volume of the magnetic field of a wave to the energy density per unit volume of its electric field in a conducting medium.

3.4. An ideal conductor with its plane face perpendicular to the Z-axis moves in the positive direction of the Z-axis with a velocity v. A plane electromagnetic wave propagates in the same direction, its electric field vector being collinear with the Y-axis [$E_x = 0$, $E_y = E_0 \cos\omega(t - z/c)$, $E_z = 0$]. Find the electric field strength of the wave formed as a result of superposition of the incident and reflected waves ($v \ll c$).

3.5. Natural light is incident on a glass surface ($n = 1.5$) at the Brewster angle. Find the degree of polarization of the refracted ray [see (16.45)].

3.6. Find the degree of polarization of a narrow beam of natural light after passing through a glass slab ($n = 1.5$) if the angle of incidence at the slab surface is equal to the Brewster angle.

3.7. The angle between the plane of oscillations of the electric field vector of an electromagnetic wave incident at the interface between the electromagnetic wave and the plane of incidence is called the oscillation azimuth. Find the oscillation azimuths α_{refl} and α_{refr} for the reflected and refracted waves if the oscillation azimuth for the incident wave is α_{in} and the angle of incidence is θ_{in}.

3.8. Find the relative deviation of the group velocity from the phase velocity in a medium with $n = 1.5$, for which $\mathrm{d}n/\mathrm{d}\lambda = 3 \times 10^4\ \mathrm{m}^{-1}$ for $\lambda = 0.5\ \mu\mathrm{m}$.

3.9. A linearly polarized wave whose electric field vector is perpendicular to the plane of incidence falls at an angle of 40° on glass ($n = 1.5$) from air. Find the reflection and transmission coefficients.

3.10. A linearly polarized wave whose electric vector oscillates in the plane of incidence falls on the surface of glass ($n = 1.6$) from air at an angle of 25°. Find the reflection and transmission coefficients.

3.11. A linearly polarized wave whose electric vector forms an angle of 30° with the plane of incidence falls on the glass surface ($n = 1.65$) at an angle of 35°. Find the reflection and transmission coefficients.

3.12. Natural light falls from glass with $n = 1.65$ at an angle of 40° on the interface with a solution whose refractive index n_2 depends on the concentration of the solution and can be varied over a wide interval. For what value of the refractive index n_2 is the reflected light linearly polarized and what is the coefficient of reflection in this case?

3.13. A light wave is incident from glass ($n = 1.55$) at an angle of 60° at its interface with air. Find the critical angle and the phase shifts of the oscillations of the electric field strength $\varphi_\perp$ and $\varphi_\parallel$ [see (17.19)].

3.14. Find the range of angles of incidence of a linearly polarized wave from air at the surface of water ($n = 1.33$) for which the reflection coefficient is more than 0.5. The plane of oscillation of the electric field vector is perpendicular to the plane of incidence.

3.15. Natural light is incident at the Brewster angle on a glass plate ($n = 1.65$). Find the reflection coefficient.

3.16. Derive the formula for the phase shift of oscillations for the electric field component of the wave perpendicular to the plane of incidence and the component lying in the plane of incidence upon total reflection.

3.17. Using the conditions described in Example 19.1, determine the phase shift between the oscillations of the electric intensity and the magnetic induction of a wave.

3.18. For what value of electric conductivity is the attenuation in a medium ($n = 1.5$) equal to 1 dB/km?

3.19. At 20 °C, the refractive index of benzene for a wavelength of 589 nm is 1.5, and its density is 0.879 g/cm^3. Using the Clausius-Mosotti equation, calculate the refractive index of benzene vapours at the boiling point (80 °C).

ANSWERS

3.1. $[k^2/(2\mu\omega^2)]\,|E_0(0)|^2 \exp(-2sz)$. 3.2. $\mu_r\gamma/(2\omega\varepsilon_0)$. 3.3. $[1 + \gamma/(\omega\varepsilon)^2]^{1/2}$. 3.4. $E_x = 0$, $E_y = E_0\{\cos\omega(t - z/c) - \cos\omega[(1 - 2v/c)(t - z/c)]\}$, $E_z = 0$. 3.5. $P = [4n^2 - (1 + n^2)^2]/[4n^2 + (1 + n^2)^2] = -0.08$. 3.6. $P = [16n^4 - (1 + n^2)^4]/[(16n^4 + (1 + n^2)^4] = 0.16$. 3.7. $\tan\alpha_{\mathrm{refr}} = \cos(\theta_{\mathrm{in}} - \theta_{\mathrm{refr}})\tan\alpha_{\mathrm{in}}$, $\tan\alpha_{\mathrm{refl}} = -\cos \times (\theta_{\mathrm{in}} - \theta_{\mathrm{refr}})\tan\alpha_{\mathrm{in}}/\cos(\theta_{\mathrm{in}} + \theta_{\mathrm{refr}})$. 3.8. $(v_g - c/n)/(c/n) = (\lambda/n) \times (\mathrm{d}n/\mathrm{d}\lambda) = 10^{-2}$. 3.9. 0.08; 0.92. 3.10. 0.04; 0.96. 3.11. 0.05; 0.95. 3.12. 1.385; 0.03. 3.13. 40.18°; 50.5°; 70.2°. 3.14. $\theta_{\mathrm{in}} > 81.2°$. 3.15. $1/2\,[(1 - n^2)/(1 + n^2)]^2 = 0.107$. 3.16. $\varphi_{\parallel} - \varphi_{\perp} = \arctan \times [\cos\theta_{\mathrm{in}}\sqrt{\sin^2\theta_{\mathrm{in}} - n_{12}^2}/\sin^2\theta_{\mathrm{in}}]$. 3.17. $\varphi = \arctan s/k = \arctan \times [\gamma/(2\varepsilon_0\omega n^2)] = 2 \times 10^{-7}$. 3.18. 0.92×10^{-6} S/m. 3.19. $n = 1 + 1.36 \times 10^{-3}$.

Chapter 4

Geometrical Optics

Basic idea:
In the optical range, the propagation of waves can be presented with a very high degree of accuracy as the transfer of wave energy along rays described by geometrical relations.

SEC. 21. GEOMETRICAL OPTICS APPROXIMATION

A method of transition from equations for waves to equations for rays is described. The concept of a ray is analyzed.

EIKONAL EQUATION. Any component of the amplitude of the light wave field in vacuum satisfies wave equation (2.12). For a medium in which the velocity of propagation of electromagnetic waves is v, the wave equation assumes the following form for any of the components:

$$\nabla^2\Phi - \frac{1}{v^2}\frac{\partial^2\Phi}{\partial t^2} = \nabla^2\Phi - \frac{n^2}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0, \qquad (21.1)$$

where we have introduced the refractive index $n = c/v$ of the medium relative to vacuum. For monochromatic waves of frequency ω, we assume that

$$\Phi(\mathrm{r}, t) = \Psi(\mathrm{r})e^{-i\omega t}. \qquad (21.2)$$

Substituting (21.2) into (21.1), we obtain the following equation for the amplitude $\Psi(\mathrm{r})$ which depends only on the coordinates:

$$\nabla^2\Psi + n^2k_0^2\Psi = 0. \qquad (21.3).$$

Here, $k_0 = \omega/c$ is the wave number corresponding to vacuum. Obviously, this quantity is connected with the wave number k in the medium through the relation

$$k = nk_0. \qquad (21.4)$$

Taking into account the relation

$$\frac{1}{\Psi}\frac{\partial^2\Psi}{\partial x^2} = \frac{\partial^2}{\partial x^2}(\ln\Psi) + \left(\frac{\partial}{\partial x}\ln\Psi\right)^2 \qquad (21.5)$$

and analogous relations for derivatives with respect to y and z, we can write Eq. (21.3) after dividing by Ψ in the following form:

$$\frac{1}{\Psi}\nabla^2\Psi + n^2k_0^2 = \nabla^2(\ln\Psi) + [\mathrm{grad}(\ln\Psi)]^2 + n^2k_0^2 = 0. \quad (21.6)$$

To solve this equaton, we put

$$\Psi(\mathrm{r}) = A(\mathrm{r})\,\mathrm{e}^{iS(\mathrm{r})}. \quad (21.7)$$

Substituting (21.7) into (21.6), we obtain

$$\nabla^2(\ln A) + [\mathrm{grad}(\ln A)]^2 - (\mathrm{grad}\,S)^2 + n^2k_0^2$$
$$+\, i[\nabla^2 S + 2\,\mathrm{grad}(\ln A)\cdot\mathrm{grad}\,S] = 0. \quad (21.8)$$

Equating the real and imaginary parts of this equation to zero, we obtain two equations for determining $A(\mathrm{r})$ and $S(\mathrm{r})$:

$$\nabla^2(\ln A) + [\mathrm{grad}(\ln A)]^2 - (\mathrm{grad}\,S)^2 + n^2k_0^2 = 0, \quad (21.9a)$$
$$\nabla^2 S + 2\,\mathrm{grad}(\ln A)\cdot\mathrm{grad}\,S = 0. \quad (21.9b)$$

These equations are considerably simplified for the optical range of wavelengths. The wave amplitude varies significantly only at distances much larger than the wavelength, i.e. at distances l of the order of the size of a lens, an optical instrument, etc., satisfying the condition

$$\lambda \ll l. \quad (21.10)$$

Comparing (21.7) with the complex form (2.39) of a plane wave, we find that

$$|\mathrm{grad}\,S|^2 \sim k^2 \sim 1/\lambda^2. \quad (21.11)$$

On the other hand, the sum of the first two terms in (21.9a) is of the order

$$\nabla^2(\ln A) + [\mathrm{grad}(\ln A)]^2 = (1/A)\nabla^2 A \sim 1/l^2 \ll 1/\lambda^2. \quad (21.12)$$

Hence these terms can be neglected in comparison with the last two terms. Equation (21.9a) can then be written as

$$\boxed{(\mathrm{grad}\,S)^2 = n^2k_0^2,} \quad (21.13)$$

called the eikonal equation.

RAY OF LIGHT. The gradient of function S is directed along the normal to surface $S = \mathrm{const}$. Hence eikonal S describes the constant-phase surfaces of a wave, while grad S leads to the concept of a ray, i.e. to the concept of propagation of light energy at a given point in a certain direction. A ray is a line the tangent to which at each point coincides with the vector grad S.

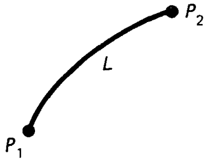


Fig. 69. Trajectory of a ray of light in a nonhomogenous medium.

The propagation of light is considered as the motion of light energy along the rays. The plane perpendicular to the light rays (i.e. plane $S = \text{const}$) is called a wave front.

RANGE OF APPLICABILITY OF THE RAY APPROXIMATION. The subject of geometrical optics is the analysis of the propagation of light in the ray approximation. It can be seen from (21.12) that this approximation is always valid when $\nabla^2 A/A$ is small in comparison with $1/\lambda^2$. In physical sense, this term describes the bending of light rays by material objects, i.e. the diffraction of light rays. Hence it can be stated that the diffraction effects are not considered in geometrical optics (see Chap. 6).

FERMAT'S PRINCIPLE. In a homogeneous medium, $S = \mathrm{k} \cdot \mathrm{r}$ ($\mathrm{k} = \text{const}$). The rays are parallel straight lines, while the wave front is a plane perpendicular to these lines. The problem is a lot more complicated for a nonhomogeneous medium.

Let points P_1 and P_2 be joined by a ray L (Fig. 69). We shall calculate the phase change along the ray. For each point of the ray we have

$$\mathrm{d}S = \operatorname{grad} S \cdot \mathrm{dr} = |\operatorname{grad} S|\,|\mathrm{dr}| = k_0 n(\mathrm{r})\,\mathrm{d}l, \tag{21.14}$$

where we have considered that dr is directed along the ray and hence coincides with grad S, and $\mathrm{d}l$ is an element of the path length. Thus, for the phase change we get the following relation

$$S = \int_{P_1}^{P_2} \mathrm{d}S = k_0 \int_{P_1}^{P_2} n(\mathrm{r})\,\mathrm{d}l, \tag{21.15}$$

where path P_1P_2 coincides with the ray.

Taking into consideration the definition of a ray and the wave front, we conclude from (21.15) that optical paths between two points of a wave front are identical at two different instants of time. If points P_1 and P_2 are joined by a curve other than the ray, all the points of the curve being in the immediate neighbourhood of the ray, and if integral (21.15) is calculated, we shall naturally obtain a different value of the integral. According to Fermat's principle, integral (21.15) has a stationary value along the ray, i.e. the first variation δS relative to the neighbouring paths of integration is equal to zero. This formulation can be varied. Considering that $\mathrm{d}l/v = \mathrm{d}t$ is the time in which the path $\mathrm{d}l$ is traversed at a velocity v, and taking into account the relation $n(\mathrm{r}) = c/v(\mathrm{r})$, we can write Eq. (21.15) in the form

$$S = k_0 c \int_{P_1}^{P_2} \frac{\mathrm{d}l}{v} = \omega \int_{P_1}^{P_2} \frac{\mathrm{d}l}{v}, \tag{21.16}$$

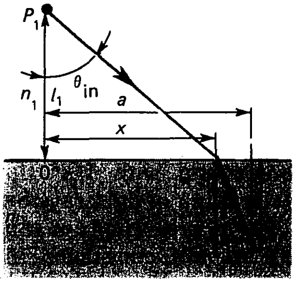


Fig. 70. To the derivation of the law of refraction with the help of Fermat's principle.

where the integral

$$\int_{P_1}^{P_2} \frac{dl}{v} \tag{21.17}$$

defines the time spent by light in covering the path from point P_1 to point P_2. Hence the Fermat principle can be formulated as follows: a ray connecting two points is the path which will be traversed in the least (stationary) time.

This principle was discovered by Fermat (1601-1665) in 1657 as the least-time principle and formulated as follows: "Nature always follows the shortest path". However, this should not be interpreted as the statement that a ray joining two points is the path whose traversal requires less time than for traversing the adjoining paths between the same two points. Fermat's statement that nature follows the shortest path is, of course, true, but there may be several shortest paths joining two points. The formulation of the stationary time spent in traversing the path between two points indicates, on the one hand, the extremal nature of this time (maximum or minimum), but does not preclude, on the other hand, the existence of several paths with identical traversal times.

Such a situation is typical of geometrical optics while constructing images.

In geometrical optics, all rays from a point on the object pass through different paths and converge at the image point. However, they all spend the same time in traversing their path. In other words, the optical lengths of all paths joining the object point with the image point are identical. This statement is called the tautochronism principle.

DERIVATION OF THE LAW OF REFRACTION FROM FERMAT'S PRINCIPLE. To illustrate the application of Fermat's principle, let us use it to derive the law of refraction. Suppose that two points P_1 and P_2 in two homogeneous media with refractive indices n_1 and n_2 separated by a plane interface (Fig. 70) are to be joined by a ray. The ray is a straight line in each homogeneous medium. Let x be the coordinate of the point at which the ray enters the second medium from the first. The total time of propagation of light from P_1 to P_2 is obviously

$$t = (n_1/c)\sqrt{l_1^2 + x^2} + (n_2/c)\sqrt{l_2^2 + (a - x)^2}, \tag{21.18}$$

(where $v = c/n$) and depends on the variable x. The stationary condition $\partial t/\partial x = 0$ assumes the form

$$n_1 x/(c\sqrt{l_1^2 + x^2}) - n_2(a - x)/(c\sqrt{l_2^2 + (a - x)^2}) = 0. \tag{21.19}$$

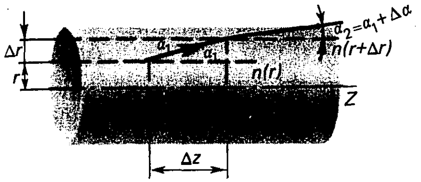


Fig. 71. To the derivation of the equation of a ray in a light guide.

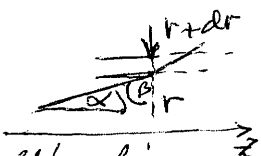
Considering that $x/\sqrt{l_1^2 + x^2} = \sin\theta_{\text{in}}$, $(a - x)/\sqrt{l_2^2 + (a - x)^2} = \sin\theta_{\text{refr}}$, we obtain the law of refraction from (21.19):

$$\sin\theta_{\text{in}}/\sin\theta_{\text{refr}} = n_2/n_1. \tag{21.20}$$

Obviously, this law is identical to Snell's refraction law (16.14).

Thus, if the law of variation of the refractive index $n(r)$ in a medium is known, we can construct a ray and thus solve the problem of propagation of light under conditions when the geometrical optics approximation is valid.

(1) see ~~Problem~~ Example 21.2 (the next page)



Snell's law gives us

$n(r)\sin\beta' = const$

(since we need to follow inhomogeneities in r) and $\alpha = \frac{\pi}{2} - \beta$

thus $n(r)\cos\alpha = const$

$tg\,\alpha = \frac{dr}{dz} \Rightarrow \frac{n(r)}{\sqrt{1+\left(\frac{dr}{dz}\right)^2}} = const$

$\Rightarrow$ (neglecting $\left(\frac{dr}{dz}\right)^2$)

$\frac{d^2r}{dz^2} = \frac{1}{n}\frac{dn}{dr}$

The rectilinear propagation of light and the phenomena of reflection and refraction were known to the ancient Greeks. The first systematic descriptions of these phenomena that have survived till the present time belong to Empedocles (490-430 B.C.) and Euclid (300 B.C.). They were aware of the law of reflection. The law of refraction of light was experimentally established in 1621 by Snell (1591-1626).

PROPAGATION OF A RAY IN A MEDIUM WITH VARYING REFRACTIVE INDEX. Let us consider the propagation of a ray in a medium in which the variation of refractive index is axially symmetric about the axis which is taken as the Z-axis (Fig. 71). The ray is assumed to be propagating in the positive direction of the Z-axis in the vicinity of the axis (paraxial ray). We denote by r the distance of the ray from the Z-axis. The Snell law of refraction in an infinitely thin layer Δr, over which the refractive index varies from $n(r)$ to $n(r + \Delta r)$, can be written as follows:

$$n(r)\cos\alpha_1 = n(r + \Delta r)\cos(\alpha_1 + \Delta\alpha). \quad ^{(1)} \tag{21.21}$$

Expanding $n(r + \Delta r)$ on the right-hand side of this equation into a Taylor series in Δr, confining to the linear term in Δr and using the trigonometric formula for the cosine of the sum of two angles, we obtain

$$n(r)\cos\alpha_1 = [n(r) + \Delta r\,\partial n/\partial r](\cos\alpha_1\cos\Delta\alpha - \sin\alpha_1\sin\Delta\alpha). \tag{21.22}$$

In the paraxial approximation, we can assume that

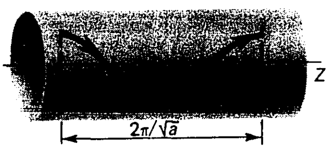


Fig. 72. Trajectory of a ray in a light guide.

$\sin\Delta\alpha \approx \Delta\alpha$ and $\cos\Delta\alpha \simeq 1$. We then obtain the following relation from (21.22) to within first-order terms in $\Delta\alpha$:

$$\frac{\partial n}{\partial r} = n(r)\frac{\Delta\alpha}{\Delta r}\tan\alpha_1, \tag{21.23}$$

where we have neglected terms with $\Delta\alpha\Delta r$, having the second order of smallness. Since $\tan\alpha_1 = \Delta r/\Delta z$, we can write in the paraxial approximation

$$\tan\alpha_1\frac{\Delta\alpha}{\Delta r} = \frac{\Delta r}{\Delta z}\frac{\Delta\alpha}{\Delta r} = \frac{\Delta\alpha}{\Delta z} = \frac{\Delta}{\Delta z}\left(\frac{\Delta r}{\Delta z}\right) \to \frac{\mathrm{d}^2 r}{\mathrm{d}z^2}. \tag{21.24}$$

Taking this relation into account, we obtain from (21.23) the equation for the ray propagation:

$$\frac{\mathrm{d}^2 r}{\mathrm{d}z^2} = \frac{1}{n(r)}\frac{\mathrm{d}n(r)}{\mathrm{d}r}. \tag{21.25}$$

Example 21.1. Analyze the propagation of a ray of light in a thin dielectric fibre (optical fibre) whose refractive index varies according to the law $n(r) = n_0(1 - ar^2/2)$. It is assumed that $ar^2/2 \ll 1\,(a > 0)$ over the entire cross section of the fibre.

Up to an accuracy of linear terms in r, Eq. (21.25) assumes the form

$$\mathrm{d}^2 r/\mathrm{d}z^2 = -ar. \tag{21.26}$$

The general solution of this equation is given by

$$r(z) = A_1\cos(\sqrt{a}z) + A_2\sin(\sqrt{a}z). \tag{21.27}$$

The values of constants A_1 and A_2 are determined from the initial conditions. The period of oscillations described by (21.27) in the direction of the Z-axis is $l = 2\pi/\sqrt{a}$ (Fig. 72). Optical fibres are widely used to control the motion of light beams. They act as light guides. If a fibre is bent and the radius of curvature is not too small (of the order of the optical wavelength), the light beam follows the bend in the fibre. A great advantage of the optical fibres is the small energy losses during propagation of light beams through them. These losses are considerably lower than the losses in a conductor when the same amount of energy is transmitted with the help of alternating currents. Hence it is advantageous to use optical fibres for transmission of information. However, the main advantage of using light for transfer of information is associated with the high frequency of light which enables a very large body of information to be imparted to a light beam. A light guide as thin as the human hair can ensure transmission of information equivalent to several hundred

!

The concept of a ray can be applied for very short wavelengths. In this case, the wave front is the surface perpendicular to the pencil of rays.

Fermat's principle does not state that the time spent by a wave in travelling along a ray is the least time, but that this time is stationary.

?

Describe any situation in which the time of wave motion along a ray from one point to another will not be the least time, but will be stationary.

telephone cables. Other significant advantages of light guides are that they have a small diameter and are made of dielectric materials which are noncorrosive and resistant to other harmful effects. Besides, the technology of their manufacture is well known.

Example 21.2. Equation (21.25) is valid only in the paraxial approximation, when the angle α and the deviation of the ray from the Z-axis are small. Analyze the problem of propagation of a ray without these restrictions, using directly the law of refraction (21.20) if $n = n_0\sqrt{1 - r^2/a^2}$. Denote the angle between the r-axis, perpendicular to the Z-axis, and the ray by $\beta = \beta(r, z)$. For $z = 0$, the ray intersects the Z-axis at an angle $\beta_0 = \beta(0, 0)$.

The law of refraction (21.20) in the form $n_1 \sin\beta_1 = n_2 \sin\beta_2$ shows that $n(r)\sin[\beta(r, z)] = \text{const}$ during the propagation of the ray. Taking into account the initial conditions, we obtain an equation for determining the ray $n \sin\beta = n_0 \sin\beta_0 = \text{const}$. Taking into account $n = n_0\sqrt{1 - r^2/a^2}$, this equation can be written in the form

$$\sqrt{1 - r^2/a^2}\,\sin[\beta(r, z)] = n_0 \sin\beta_0\,. \tag{21.28}$$

It can be seen that $|r| < |a|$, i.e. the ray cannot depart from the Z-axis beyond a distance $|a|$. It can be shown that the equation for the ray is given by the formula $r = a\cos\beta_0 \sin[z/(a\sin\beta_0)]$. This, however, is not a simple problem.

SEC. 22. LENSES, MIRRORS AND OPTICAL SYSTEMS

The rules for describing a ray in the paraxial approximation by the matrix methods are introduced.

PARAXIAL APPROXIMATION. Let us consider the passage of rays through a lens without imposing any restrictions on its thickness. We assume that the Z-axis coincides with the axis of the lens (Fig. 73) and that the light ray which propagates along positive direction of the Z-axis lies in the XZ-plane. The lens has spherical surfaces with radii of curvature r_1 and r_2. The quantities pertaining to the spherical surface encountered first by the rays are denoted by subscript 1, and the subscript 2 is used for the second surface. The material of the lens has a refractive index n_2, while n_1 is the refractive index of the medium in which the lens is situated. Calculations involve small angles which are equal in the order of magnitude to the angles between the rays and the Z-axis. All calculations are carried out in the paraxial approximation, i.e.

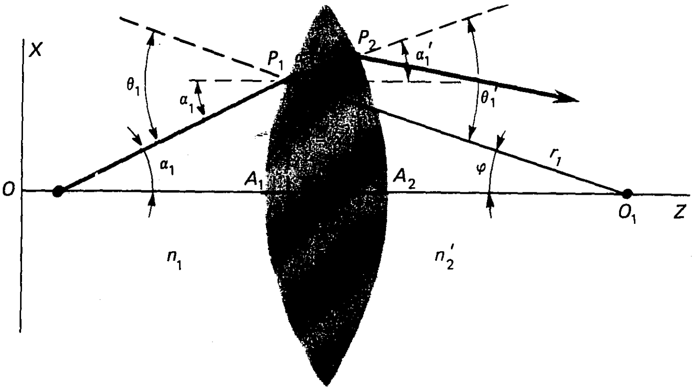


Fig. 73. To the calculation of the passage of a ray through a lens.

in the assumption of a close proximity of the rays to the Z-axis. Taking into account the expansions

$$\begin{aligned}\sin\alpha &= \alpha - \alpha^3/(3!) + \alpha^5/(5!) - \dots,\\ \cos\alpha &= 1 - \alpha^2/(2!) + \alpha^4/(4!) - \dots,\\ \tan\alpha &= \alpha + \alpha^3/3 + 2\alpha^5/15 + \dots,\end{aligned} \tag{22.1}$$

we conclude that for small angles α, we can put $\sin\alpha \approx \alpha$, $\cos\alpha \simeq 1$, and $\tan\alpha \approx \alpha$. For example, if the angle is equal to 6°, then $\alpha = 2\pi(6/360) \simeq 0.1$ and hence $\sin\alpha \approx \alpha$ to within 10^{-4}, $\tan\alpha \approx \alpha$ to within 10^{-3}, and $\cos\alpha \simeq 1$ to within 10^{-2}. Hence the paraxial approximation gives quite accurate results over a fairly wide range of conditions.

REFRACTION AT A SPHERICAL SURFACE. In the paraxial approximation, Snell's law of refraction at point P_1 has the form

$$n_1\theta_1 = n_2\theta_1'. \tag{22.2a}$$

The angles of inclination of the incident and refracted rays at point P_1, viz. α_1 and α_1', satisfy the relations

$$\theta_1 = \alpha_1 + \varphi, \tag{22.2b}$$

$$\theta_1' = \alpha_1' + \varphi, \tag{22.2c}$$

which can be used to present (22.2a) in the form

$$n_1(\alpha_1 + \varphi) = n_1'(\alpha_1' + \varphi). \tag{22.3}$$

To preserve uniform notation, we have put in this equation $n_1' = n_2$. Considering that $\varphi = x_1/r_1$ (paraxial approximation), we obtain instead of (22.3)

$$n_1(\alpha_1 + x_1/r_1) = n_1'(\alpha_1' + x_1/r_1), \tag{22.4}$$

or

$$n_1'\alpha_1' = (n_1 - n_1')(x_1/r_1) + n_1\alpha_1. \tag{22.5}$$

The last equation can be used to determine the angle of inclination α_1' of the refracted ray if the angle of inclination α_1 of the incident ray and the distance x_1 between the axis and the point of incidence are known. Obviously, the distance x_1' for the refracted ray at point P_1 is equal to x_1, i.e. $x_1' = x_1$.

MATRIX NOTATION. For simplicity of notation, it is convenient to present formula (22.5) together with the equality $x_1' = x_1$ in the matrix form:

$$\begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix} = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix}, \tag{22.6}$$

where

$$k_1 = (n_1 - n_1')/r_1, \tag{22.7}$$

is the refractive power of surface 1. The two-row matrix

$$R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \tag{22.8}$$

on the right-hand side of (22.6) is called the refraction matrix of surface 1.

PROPAGATION OF A RAY IN A LENS. The refracted ray intersects the second surface at a distance

$$x_2 = x_1' + \Delta \tan \alpha_1' = x_1' + \Delta\alpha_1' \tag{22.9}$$

from the axis. In the paraxial approximation, we can assume that Δ is equal to the thickness A_1A_2 of the lens, while the angle α_1' is the angle of inclination α_2 of the ray in the second medium ($\alpha_1' = \alpha_2$ simply indicates a change in notation). Consequently, in view of the fact that $n_1' = n_2$ and $\alpha_1' = \alpha_2$, Eq. (22.9) can be written in the matrix form as follows:

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \dfrac{\Delta_1'}{n_1'} & 1 \end{pmatrix} \begin{pmatrix} n_1'\alpha_1' \\ x_1' \end{pmatrix}, \tag{22.10}$$

while the two-row matrix

$$T_{21} = \begin{pmatrix} 1 & 0 \\ \Delta_1'/n_1' & 1 \end{pmatrix} \tag{22.11}$$

is called the transfer matrix T_{21}. It describes the propagation of the ray from the first to the second refracting surface inside the lens. It should be noted that the quantities $n_2\alpha_2$ and x_2 characterize the ray at a point on the second surface before refraction.

REFRACTION OF THE RAY AT THE SECOND SPHERICAL SURFACE. This is described in the same way as (22.6) with the help of the refraction matrix for the second surface:

$$R_2 = \begin{pmatrix} 1 & -k_2 \\ 0 & 1 \end{pmatrix}, \tag{22.12}$$

where

$$k_2 = (n_2' - n_2)/r_2. \tag{22.13}$$

Here, r_2 is taken with the minus sign, since the centre of curvature of this surface lies on the negative side of the Z-axis with respect to the surface (for the incident ray, this surface is concave and not convex). Consequently, we can write the following matrix equation for refraction at the second surface:

$$\begin{pmatrix} n_2'\alpha_2' \\ x_2' \end{pmatrix} = R_2 \begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix}. \tag{22.14}$$

REFRACTION OF A RAY BY A LENS. Combining (22.6), (22.10) and (22.14), we obtain the following relation between the characteristics of the ray at the entrance to the lens and at the exit from it:

$$\begin{pmatrix} n_2'\alpha_2' \\ x_2' \end{pmatrix} = R_2 T_{21} R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix}. \tag{22.15}$$

!

The matrix method can be used to trace the path of the rays propagating through an optical system. This method is quite suitable for computer calculations.

?

Write down the refraction matrix for a spherical surface. How is the transfer matrix constructed?
How is the matrix describing the reflection at a spherical surface obtained?

PROPAGATION OF A RAY THROUGH AN OPTICAL SYSTEM. The propagation of a ray to the right of the lens is described by the transfer matrix T_{23} which is constructed in the same way as (22.11), but now Δ_1' is replaced by Δ_2', the distance from point A_2 to a point of intersection of the plane in which we want to determine the parameters of the ray with the Z-axis (Fig. 73). If the ray encounters another lens on its path, the refraction at the first surface of this lens is described by the refraction matrix of this surface, and so on. Thus, the calculations of the propagation of a ray through an optical system is reduced to the multiplication of matrices describing the refractive power of the lens surface and the transfer matrices. The sign convention must be borne in mind during such multiplication: if the refracting surface encountered by the ray is convex, its radius of curvature must be taken with the positive sign, while negative sign must be used for a concave surface; the angles α measured from the Z-axis in the anticlockwise direction are positive while those measured in the clockwise direction are negative. The distances measured from the Z-axis from left to right are positive, while those measured from right to left are negative. The distances measured above the Z-axis are positive while those below the axis are negative. It should be also recalled that the thickness A_1A_2 of the lens at its axis is taken for the quantity Δ.

REFLECTION FROM SPHERICAL SURFACES. Reflection from spherical surfaces is considered as refraction into a medium with a negative refractive index $-n$, where n is the refractive index of the medium from which the ray is incident on the reflecting surface. For the rest, the matrix describing reflection is absolutely the same as the matrix describing refraction. For example, the reflection from the concave surface of a spherical mirror with a radius of curvature r_2 into a medium with refractive index n_2 is described by a matrix of the type (22.12) with $n_2' = -n_2$, i.e.

$$R_2 = \begin{pmatrix} 1 & -(-n_2 - n_2)/r_2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2n_2/r_2 \\ 0 & 1 \end{pmatrix}. \tag{22.16}$$

Since the centre of curvature of the spherical surface lies towards the negative values of z, $r_2 = -|r_2|$. After reflection, the ray moves in the negative direction of the Z-axis. This is taken into account in the next transfer matrix by using minus signs for the path length and the refractive index. If a refractive surface is encountered in the path of the ray, the refractive index of the medium from which the ray is incident on this surface is considered negative as before, the refractive index of the medium into which the refracted ray enters is taken with the minus sign (since the ray after refraction continues to move in the negative direction of the Z-axis), while the sign of the radius of curvature of the surface is determined in the usual manner. For example, if after reflection from a concave surface, described by matrix (22.16), a ray of light moving in the direction of negative values of z encounters a concave surface with radius of curvature r_1 (the surface is concave if we look in the negative direction of the Z-axis), then r_1 is taken with the plus sign according to the common rule, since its centre of curvature lies in the positive direction of the Z-axis from the surface. If n_2 is the refractive index of the second medium, the refraction matrix has the form

$$R_3 = \begin{pmatrix} 1 & -\dfrac{-n_1 - (-n_2)}{r_1} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \dfrac{n_1 - n_2}{r_1} \\ 0 & 1 \end{pmatrix}. \tag{22.17}$$

Thus, the reflection from mirrors is analyzed by the matrix methods in the same way as refraction. Only the signs of the quantities appearing in the reflection matrix should be carefully chosen. The optical systems containing mirrors are considered in accordance with the general rules of matrix methods.

Example 22.1. One of the spherical surfaces of a double-convex lens is silvered and reflecting. For definiteness, we assume that the right spherical surface with radius of curvature r_2 of the lens shown in Fig. 75 is silvered. The lens is in air ($n = 1$) and the refractive index of the material of the lens is $n_2 > 1$. The radii of curvature of the surfaces of the lens are r_1 and r_2 (in accordance with the general rule, r_2 is a negative quantity, i.e. $r_2 = -|r_2|$). The ray of light is incident from the left. Find the transfer matrix from the entrance of the ray into the lens to its exit from the same surface.

In accordance with the rules described above, the refraction matrix at the first surface has the form

$$R_1 = \begin{pmatrix} 1 & -(n_2 - 1)/r_1 \\ 0 & 1 \end{pmatrix}. \tag{22.18}$$

The motion of the ray after refraction at the first surface up to its arrival at the reflecting surface is described by the matrix

$$T_{21} = \begin{pmatrix} 1 & 0 \\ \Delta_1'/n_2 & 1 \end{pmatrix}. \tag{22.19}$$

The backward reflection from the second surface into the lens is described by matrix (22.16). After reflection, the motion of the ray up to its arrival at the first surface (the second passage of the ray is denoted by the subscript 3) is described by the matrix

$$T_{32} = \begin{pmatrix} 1 & 0 \\ -\Delta_1'/(-n_2) & 1 \end{pmatrix}. \tag{22.20}$$

The refraction matrix at the first surface for the motion of the ray in the negative direction is given by expression (22.17) with $n_1 = 1$, i.e.

$$R_3 = \begin{pmatrix} 1 & -(n_2 - 1)/r_1 \\ 0 & 1 \end{pmatrix}. \tag{22.21}$$

Consequently, the required transfer matrix is

$$S_{31} = R_3 T_{32} R_2 T_{21} R_1 . \tag{22.22}$$

Multiplying the matrices, we obtain all the characteristics of the optical system in accordance with the general rules.

SEC. 23. OPTICAL IMAGE

The matrix approach towards constructing images in optical systems is described.

MATRIX OF AN OPTICAL SYSTEM. Equation (22.15) can be presented in the following form:

$$\begin{pmatrix} n_2'\alpha_2' \\ x_2' \end{pmatrix} = S_{21}\begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix}; \tag{23.1}$$

$$S_{21} = R_2 T_{21} R_1 = \begin{pmatrix} 1 & -k_2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \Delta_1'/n_1' & 1 \end{pmatrix}\begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} b & -a \\ -d & c \end{pmatrix}; \tag{23.2}$$

$$a = k_1 + k_2 - k_1 k_2 \Delta_1'/n_1', \tag{23.3a}$$

$$b = 1 - k_2 \Delta_1'/n_1', \tag{23.3b}$$

$$c = 1 - k_1 \Delta_1'/n_1', \tag{23.3c}$$

$$d = -\Delta_1'/n_1', \tag{23.3d}$$

where a, b, c and d are Gauss' constants. Not all these constants are independent of one another. A direct verification shows that the determinants of R_1, R_2 and T_{21} are equal to unity, and hence

$$\det S_{21} = bc - ad = \det(R_2 T_{21} R_1) = 1. \tag{23.4}$$

Only three of the four Gauss constants are independent. Usually, a, b and c are chosen as these independent constants.

TRANSFORMATION OF A RAY FROM OBJECT PLANE TO IMAGE PLANE. The parameters of the ray $(n_1\alpha_1, x)$ are given in a plane at a distance l to the left of A_1 (object plane). The ray is characterized by parameters $(n_2'\alpha_2', x')$ in a plane at a distance l' to the right of A_2 (image plane). According to the general rule described in Sec. 22, these parameters are connected through the following relation:

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l'/n_2' & 1 \end{pmatrix}\begin{pmatrix} b & -a \\ -d & c \end{pmatrix}\begin{pmatrix} 0 & 0 \\ -l/n_1 & 1 \end{pmatrix}\begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix}, \tag{23.5}$$

where the distance to the object plane is measured to the left of the lens and hence appears with the negative sign in the last square matrix in (23.5). The matrices in (23.5) can be multiplied together and the formula assumes the form

$$\begin{pmatrix} n_2'\alpha_2' \\ x' \end{pmatrix} = Q_{21}\begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix}, \tag{23.6}$$

where

$$Q_{21} = \begin{pmatrix} b + al/n_1 & -a \\ bl'/n_2' - d + al'l/(n_2'n_1) - cl/n_1 & c - al'/n_2' \end{pmatrix} \tag{23.7}$$

is the matrix of transformation from object to image.

It can be seen from (23.6) that the increase in $M = x'/x$ depends on α_1. In other words, rays emerging at different angles from one point on the object do not converge at one point on the image plane. Consequently, a point on the object will be represented on the image plane in the form of a smeared spot. By an image we mean a transformation of the object plane into the image plane such that all the rays emanating from a point on the object converge at one point on the image plane and all the points are transformed with equal magnification. For this purpose, we must make the transformation independent of the angle α_1. This is attained by putting the term responsible for such a dependence in matrix (23.7) equal to zero:

$$bl'/n_2' - d + al'l/(n_2'n_1) - cl/n_1 = 0. \tag{23.8}$$

The magnification is written as

$$M = x'/x = c - al'/n_2'. \tag{23.9}$$

Since the determinants of all the factors comprising the matrix Q_{21} are equal to unity, we have $\det Q_{21} = 1$. In this case, taking into account (23.9) and the validity of condition (23.8), we obtain

$$b + al/n_1 = 1/M \tag{23.10}$$

and hence

$$Q_{21} = \begin{pmatrix} 1/M & -a \\ 0 & M \end{pmatrix}. \tag{23.11}$$

CARDINAL ELEMENTS OF AN OPTICAL SYSTEM. Let us find the planes for whose points the magnification $M = 1$. In accordance with (23.10), the object plane H is situated at a distance

$$l_H = n_1(1 - b)/a, \tag{23.12}$$

from the point A_1 (Fig. 74), while the image plane H' is situated, in accordance with (23.9), at a distance (Fig. 74)

$$l_H' = n_2'(c - 1)/a, \tag{23.13}$$

from A_2.

The planes H and H' are called the principal planes, and

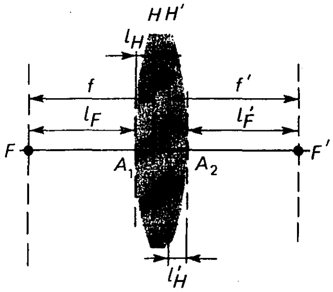


Fig. 74. Cardinal elements of an optical system.

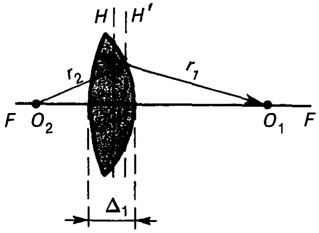


Fig. 75. An example showing the arrangement of principal planes inside a lens.

their intersection with the axis of the system coinciding with the Z-axis occurs at the principal points of the system.

Let us find a point on the axis of the system, at which the rays that were parallel to the axis before passing through the optical system converge. For $M = 0$, we obtain from (23.9) the following equation for determining this point:

$$c - al'_F/n'_2 = 0, \tag{23.14}$$

from which it follows that

$$l'_F = n'_2 c/a. \tag{23.15}$$

This point is called the focus of the optical system. Its distance from the principal plane H' is called the focal length (Fig. 74):

$$\boxed{f' = l'_F - l'_H = n'_2/a.} \tag{23.16}$$

In the same way, we can determine the point the rays from which become parallel to the optical axis after passing through the system. For $M = \infty$, we obtain from (23.10) the following equation for this point:

$$l_F = -n_1 b/a. \tag{23.17}$$

This point is also called the focus of the optical system and its distance from the principal plane H is also called the focal length (Fig. 74):

$$f = l_F - l_H = -n_1/a. \tag{23.18}$$

The planes normal to the optical axes and passing through the foci are called the focal planes of the optical system. The principal planes and the focal planes of an optical system are its cardinal elements. If the cardinal elements are known, we can determine the action of the optical system on the rays entering it without knowing the details of the passage of the rays inside the system.

PHYSICAL MEANING OF THE GAUSS CONSTANTS. For air, we can assume that $n_1 = n'_2 = 1$. In this case, relations (23.16) and (23.18) lead to the equalities

$$-f = f' = 1/a, \tag{23.19}$$

i.e. a is the reciprocal of the focal length. Taking (23.19) into account, we obtain from (23.17)

$$b = l_F/f \tag{23.20}$$

and, similarly,

$$c = l'_F/f'. \tag{23.21}$$

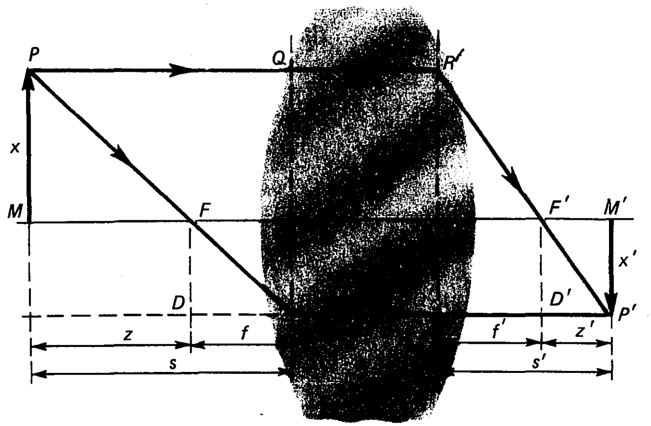


Fig. 76. Image construction from the cardinal elements of an optical system.

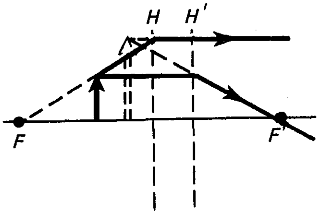


Fig. 77. Virtual image formation.

Thus, *b* and *c* characterize the mutual arrangement of the principal and focal planes. For example, if the principal plane *H'* lies inside the lens, the quantity *c* defines the fraction of the focal length lying outside the lens (Fig. 75). It should be noted that the principal planes of a fairly thick converging lens lie within the lens and are asymmetric with respect to the centre if the radii of curvature $|r_1|$ and $|r_2|$ of the refracting surfaces are different.

CONSTRUCTION OF IMAGES. The image of a point is constructed from two rays, one of which is parallel to the optical axis and the other passes through the focus. The first ray is parallel to the optical axis until it intersects the principal plane *H'* (Fig. 76), behind which it passes through the focus *F'*. The second ray is drawn up to its intersection with the principal plane *H* behind which it is parallel to the optical axis until it intersects the first ray. The point of intersection of these two rays is the image of the object point. If the distance between the object and the principal plane is less than the focal length, then instead of the second ray its extension in the backward direction passes through the focus. A pair of divergent rays emerges from the system (Fig. 77). The intersection of these rays forms a virtual image of the object point. All the remaining cases of image formation can be reduced to these two cases.

LENS EQUATION. The similarity of triangles *RDF*, *RQP* and *FMP* (Fig. 76) yields

$$(x + x')/s = x'/f = x/z, \qquad (23.22)$$

while the similarity of triangles *P'D'F'*, *F'H'R'* and *P'Q'R'* leads to the relations

$$(x + x')/s' = x/f' = x'/z'. \qquad (23.23)$$

These relations give the expression

$$x = zx'/f = x'f'/z'. \tag{23.24}$$

The last equality in (23.24) gives

$$zz' = ff'. \tag{23.25}$$

This equality is called the lens equation in Newton's form. From (23.22) and (23.23), we obtain

$$(x + x')/x' = s/f = s'/z' = s'/(s' - f'), \tag{23.26}$$

where $z' = s' - f'$. The equality $s/f = s'/(s' - f')$, written in the form

$$f'/s' + f/s = 1, \tag{23.27}$$

is called the lens equation in Gaussian form and is equivalent to (23.25). It follows from (23.22) and (23.23) that

$$x + x' = sx'/f = s'x/f. \tag{23.28}$$

The last of these equalities gives the magnification of a lens:

$$M = x'/x = s'f/(sf'). \tag{23.29}$$

THIN LENSES. Taking into account (23.16) and (23.18), we obtain the following relations from (23.3a):

$$a = n_1/f = n_2'/f' = k_1 + k_2 - k_1k_2\Delta_1'/n_1'. \tag{23.30}$$

Using the values of k_1 and k_2 given by (22.7) and (22.13), and assuming that the lens is in air ($n_1 = n_2' = 1$), we can transform (23.30)

$$\frac{1}{f'} = (n_1' - 1)\left[\frac{1}{r_1} - \frac{1}{r_2} + \frac{(n_1' - 1)\Delta_1'}{n_1'r_1r_2}\right]. \tag{23.31}$$

This formula expresses the focal length of a lens in terms of the refractive index of its material and its geometrical parameters.

By definition, a thin lens is one for which the third term in the brackets on the right-hand side of (23.31) can be neglected. For a thin lens the equation thus assumes the following form:

$$\frac{1}{f'} = (n_1' - 1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right). \tag{23.32}$$

!

The matrix for an optical system is obtained by multiplying matrices describing the passage of a ray through the composite parts of the optical system.

For thin lenses, the third term in the brackets on the right-hand side of (23.31) must be much smaller than either of the first two terms. Since n_1' is of the order of unity, we conclude that the thickness Δ_1' of the lens must be much smaller than each of the radii of curvature $|r_1|$ and $|r_2|$ of the lens surfaces, i.e.

$$\Delta_1' \ll |r_1|, \quad \Delta_1' \ll |r_2|. \tag{23.33}$$

The conditions (23.3b) and (23.3c) then assume the form

$$b = 1, \quad c = 1, \tag{23.34}$$

while Eqs. (23.12) and (23.13) are reduced to the relations

$$l_H = l_H' = 0, \tag{23.35}$$

i.e. a thin lens is assumed to have zero thickness, and both principal planes coincide with it. The focal length of such a lens is equal to the distance between the lens and the focus, and matrix (23.2) becomes

$$\mathrm{S}_{21} = \begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix}, \tag{23.36a}$$

where $a = 1/f'$. For a thin lens, Eq. (23.27) is simplified:

$$\frac{1}{s'} + \frac{1}{s} = \frac{1}{f'}. \tag{23.36b}$$

The reciprocal of the focal length of a lens is the quantity

$$\Phi = 1/f', \tag{23.36c}$$

called the optical power of the lens. It is expressed in diopters. A diopter is the optical power of a lens with a focal length 1 m.

A SYSTEM OF THIN LENSES. Matrix (23.36a) transforms the parameters of a ray entering a thin lens into the parameters of the ray emerging from the lens. The propagation of a ray up to a thin lens and behind it is described by the transfer matrix T of the type (22.11). If, for example, we have two lenses, the matrix of transformation of the ray from its entrance into the first lens up to its exit from the second lens is equal to the product of matrices describing the transformation of the ray in the lenses and matrix T describing the propagation of the ray between the lenses, taken from right to left in the same order in which the ray propagates in the system of lenses. The matrices S and T can be assigned subscripts indicating the planes of a given matrix between which the ray undergoes a transformation. It is expedient to assign different subscripts to different sides of a thin lens. For example, denoting by A_1 and A_2 the planes of a thin lens through which the ray enters and leaves

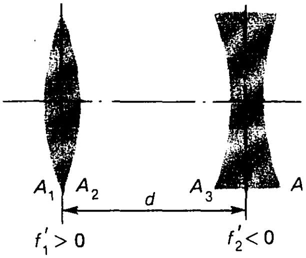


Fig. 78. To the determination of the transfer matrix for a system of two lenses.

the lens, we can write matrix (23.36) in the form

$$S_{21} = {}_{A_2}\begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix}_{A_1}. \tag{23.37}$$

Similar notation is also used for the transfer matrix T.

Let us suppose that an optical system is composed of two lenses with the focal lengths $f_1' > 0$ and $f_2' < 0$ (convergent and divergent lenses) and that the separation between the two lenses is d (Fig. 78). The lenses are shown in the figure by dashed lines to indicate that their thickness is neglected. The transfer matrix S_{41} transforming the parameters of the ray from its entrance into the first lens on plane A_1 to its exit from the second lens on plane A_4 can be written, on the basis of (23.37) and (22.11), in the following form:

$$S_{41} = {}_{A_4}\begin{pmatrix} 1 & -1/f_2' \\ 0 & 1 \end{pmatrix}_{A_3}\ {}_{A_3}\begin{pmatrix} 1 & 0 \\ d & 1 \end{pmatrix}_{A_2}\ {}_{A_2}\begin{pmatrix} 1 & -1/f_1' \\ 0 & 1 \end{pmatrix}_{A_1}, \tag{23.38}$$

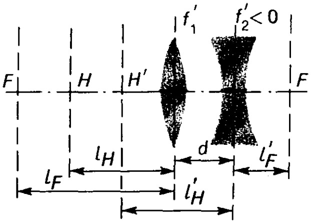


Fig. 79. Construction of cardinal planes of a system formed by a convergent lens and a divergent lens.

where it is assumed that the refractive index of the medium between the lenses is equal to unity. In order to construct the image formed by this lens system, we must multiply the matrices in (23.38) and find the values of the Gauss constants a, b, c and d, defined in (23.2):

$$S_{41} = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix}. \tag{23.39}$$

Knowing the Gauss constants, we can calculate l_H, l_H', l_F and l_F' from formulas (23.12), (23.13), (23.15) and (23.17). We draw the cardinal planes of the system and the image in accordance with the general rules considered for Fig. 76. Figure 79 shows the cardinal planes of the system shown in Fig. 78. It should be recalled that for negative values of l_H, l_H', l_F and l_F', these planes are drawn to the left of the reference point, while for the positive values of these quantities, the planes are drawn to the right. In Fig. 79, l_F, l_H and l_H' are assumed to be negative, while l_F' is assumed to be positive.

UTILIZATION OF COMPUTERS. The matrix representation of the transformations suffered by a ray in an optical system makes it very convenient to use computers for the analysis of optical systems, since standard computer programmes exist for matrices. Hence the design and calculation of optical systems are made with the help of computers, although analytical and graphic methods continue to be quite effective means for analysis of general laws.

Example 23.1. Calculate the elements of a double-convex lens for which $r_2 = -2$, $r_1 = +3$, $n_1' = 1.5$ and $\Delta_1' = 1$ (see

?

What is the physical meaning of Gauss constants and how are they used to obtain the matrix of an optical system?
Why are only three of the four Gauss constants independent?
Define the cardinal elements of an optical system.

Fig. 75). The lens is in vacuum (air). The units of length are arbitrary, hence the lengths are given in the problem only through numerical digits without any units.

In accordance with the notation used in this section, the units encountered in the calculations can be represented as follows:

$$\Delta_2 = \Delta_1' = 1;\ n_1 = 1;\ n_2 = n_1' = 1.5;\ n_2' = 1. \tag{23.40}$$

From (22.7) and (23.13), we obtain

$$k_1 = (n_1' - n_1)/r_1 = 0.17; \quad k_2 = (n_2' - n_2)/r_2 = 0.25.$$

Hence we obtain the Gauss constants with the help of formulas (23.3a-d):

$$\begin{aligned} a &= 0.17 + 0.25 - 0.17 \times 0.25 \times 1/1.5 = 0.39; \\ b &= 1 - 0.25 \times 1/1.5 = 0.84; \\ c &= 1 - 0.17 \times 1/1.5 = 0.87; \\ d &= -1/1.5 = -0.67. \end{aligned} \tag{23.41a}$$

The equality of the determinant of the Gauss constants to unity ($bc - ad = 0.99 \simeq 1$) can be considered as a fairly good verification of the correctness of the computations. From (23.12) and (23.13), we obtain

$$\begin{aligned} l_H &= 1\,(1 - 0.84)/0.39 = 0.41; \\ l_H' &= 1\,(0.87 - 1)/0.39 = -0.33. \end{aligned} \tag{23.41b}$$

Hence, with the help of (23.17) and (23.15), we obtain

$$\begin{aligned} l_F &= -1 \times 0.84/0.39 = -2.15; \\ l_F' &= 1 \times 0.87/0.39 = 2.23. \end{aligned}$$

It can be concluded from (23.41b) that the cardinal planes are situated within the lens, and in accordance with (23.18) and (23.16) the focal lengths are $f = -2.15 - 0.41 = -2.56$, $f' = 2.23 - (-0.34) = 2.56$. This also indicates the correctness of the calculations.

Example 23.2. Calculate a system composed of a double-convex lens and a double-concave lens (see Fig. 78). The lenses are assumed to be thin, and $d = 10$, $f_1 = 15$, $f_2 = -12$. The refractive index of the medium in which the lenses are placed is equal to unity.

To begin with, we obtain the transfer matrix in accordance with formula (23.38):

$$S_{41} = \begin{pmatrix} 1 & 1/12 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 10 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1/15 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1.83 & -0.04 \\ 10 & 0.34 \end{pmatrix}. \tag{23.42}$$

Consequently, the Gauss constants are

$$a = 0.04; \quad b = 1.83; \quad c = 0.34; \quad d = -10. \tag{23.43}$$

As in Example 23.1, we can use these constants to determine all the characteristics of the optical system. Further, we can analyze the path of the rays and construct the image graphically. However, this can be done with the help of matrix Q_{21} in accordance with formula (23.6).

SEC. 24. OPTICAL ABERRATION

Aberrations in optical systems and the methods of reducing or eliminating them are described.

SOURCES OF ABERRATION. The definition of the concept of image contains the condition that all the rays emerging from a certain point on the object must converge at one and the same point in the image plane and all the points of the object must be magnified to the same extent in the same plane.

For paraxial rays, these conditions are always observed in the image formation though not very accurately. In other words, the paraxial approximation gives an approximate, albeit quite precise, description of the paraxial rays. Hence the ideal pattern of images obtained in the paraxial approximation is not realized in actual practice. The deviations of the actual image from the ideal one are called aberrations. For paraxial rays, aberrations are small and can be neglected. If, however, the rays are not paraxial, the aberrations become significant and considerably distort the image. Hence the first cause of aberrations lies in that the lenses formed by spherical surfaces do not refract the rays in the same way as it is assumed in the paraxial approximation. For example, the foci for the rays incident on the lens at different distances from the optical axis are different, and so on. Such aberrations are called geometrical aberrations, and can be classified according to certain criteria. For example, the paraxial approximation is based on the fact that the exact formulas for the expansion of the sine into series (22.1) are truncated at the first term which is proportional to α. The term of the order of α^3, which is omitted in the paraxial approximation, leads to a third-order aberration.

The next term in the expansion describes the aberration of the fifth order, which is usually smaller than the third-order aberration, and so on.

The second source of aberrations is associated with dispersion of light. Since the refractive index depends on frequency, the focal length and other characteristics of the system also depend on frequency. Hence the rays corresponding to radiation of different frequencies and emitted by the same source do not converge at the same point on the image plane even when the rays corresponding to each frequency form an ideal image of the object. Such aberrations are called chromatic aberrations.

The analysis of geometrical aberrations boils down to the consideration of those factors that were neglected in the paraxial approximation. In principle, it is a simple but extremely cumbersome and laborious task. Hence we shall simply describe the essence of the problem without going into detail of the mathematical aspect. The same is true for chromatic aberrations. Exact calculations of these effects are carried out with the help of computers.

EXACT TRANSFORMATION MATRICES. In order to determine the exact transformation matrices of the parameters of a ray upon refraction at a spherical surface, we must carry out exact calculations instead of the approximate calculations, starting from formula (22.2) onwards (see Fig. 73). Instead of formula (22.2), we write the Snell law:

$$n_1 \sin\theta_1 = n_1' \sin\theta_1' . \tag{24.1}$$

An exact consideration (and not in the paraxial approximation) of geometrical relations between angles and lengths in Fig. 73 leads to the following relation instead of (22.5):

$$n_1' \sin\alpha_1' = (n_1 \cos\theta_1 - n_1' \cos\theta_1')(x_1/r_1) + n_1 \sin\alpha_1 . \tag{24.2}$$

In the paraxial approximation, this relation coincides with (22.5). Taking into account $x_1' = x_1$, we can present (24.2) in the matrix form in the same way as (22.6):

$$\begin{pmatrix} n_1' \sin\alpha_1' \\ x_1' \end{pmatrix} = \begin{pmatrix} 1 & -\tilde{k}_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1 \sin\alpha_1 \\ x_1' \end{pmatrix}, \tag{24.3}$$

where

$$\tilde{k}_1 = (n_1' \cos\theta_1' - n_1 \cos\theta_1)/r_1 .$$

Hence the exact refraction matrix has the form

$$\tilde{R}_1 = \begin{pmatrix} 1 & -\tilde{k}_1 \\ 0 & 1 \end{pmatrix}, \tag{24.4}$$

which is analogous to (22.8), except that the value of k_1 has been changed. For the ray after refraction but before arrival at the second refractive surface, we can write instead of (22.9)

$$x_2 = x_1' + L_1' \sin\alpha_1' , \tag{24.5}$$

where L_1' is the length of the segment $|P_1P_2|$. Further, in the same way as during the derivation of (22.10), we consider that $n_1' = n_2$, $\alpha_1' = \alpha_2$, and hence

$$n_1' \sin\alpha_1' = n_2 \sin\alpha_2 . \tag{24.6}$$

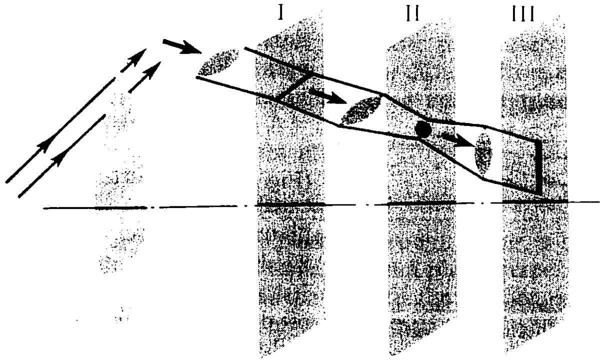


Fig. 80. Astigmatism of inclined beams.

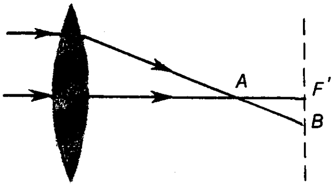


Fig. 81. Longitudinal aberration $F'A$ and transverse aberration $F'B$.

In the matrix form Eqs. (24.5) and (24.6) can be written as follows:

$$\begin{pmatrix} n_2 \sin \alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ L_1'/n_1' & 1 \end{pmatrix} \begin{pmatrix} n_1' \sin \alpha_1' \\ x_1' \end{pmatrix}. \tag{24.7}$$

Hence the matrix

$$\tilde{T}_{21} = \begin{pmatrix} 1 & 0 \\ L_1'/n_1' & 1 \end{pmatrix}, \tag{24.8}$$

which is analogous to matrix (22.11), is the transfer matrix during an exact analysis of the light propagation at any angle in the medium. In formula (24.8), the value of L_1' is not constant and must be calculated at each step from the values of α_1' and x_1' calculated at the previous step, and the parameters characterizing the second refracting surface of the lens.

The refraction at the second surface of the lens is analyzed with the help of a matrix of the type (24.4) with appropriate values of k_1, while the propagation of the ray in front of and behind the lens is described by matrices of the type (24.8). These matrices are used in the same way as in the paraxial approximation for (23.5).

SPHERICAL ABERRATION. The most serious aberrations are associated with the fact that the circular cross section of a beam incident at an angle to the lens is not conserved behind the lens (Fig. 80). The rays that are parallel to the optical axis do not converge behind the lens at a single point on the image plane (Fig. 81), and a point on the axis of the system has a coma-shaped image. Let us begin the description from the second of these aberrations, called spherical aberration. It can be interpreted as a transverse aberration or a longitudinal one (Fig. 81). If in the problem under investigation it is significant that the ray does not intersect the axis at

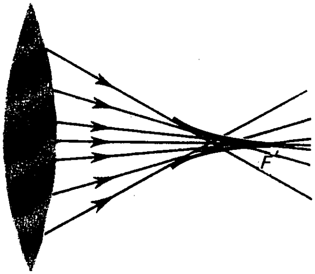


Fig. 82. Caustic surface in the case of spherical aberration.

the paraxial focus, we speak of longitudinal spherical aberration. If, however, the deviation of the ray from the axis in the paraxial focal plane is important, we speak of transverse spherical aberration.

A beam of rays parallel to the axis forms after refraction a set of cones with the vertices lying on the axis of the lens (Fig. 82). The surface enveloping this set of cones is called the caustic surface, and the cross section of this surface by any plane passing through a ray is called the caustic curve. Figure 82 shows the caustic curve in the case of spherical aberration. Its cross sections by planes perpendicular to the axis are circles of different radii. A parallel beam of rays is created by a glowing spot on the axis very far from the lens surface. Hence the bright circles play the role of the images of the point in different planes. The focus F' is determined in the paraxial approximation and plays the role of focus only for paraxial rays, i. e. for the rays that pass through the lens near its axis. The brightest and the smallest image of the point is formed by the lens in plane M' which does not pass through the paraxial focus F'. Hence, to reduce the transverse spherical aberration of a lens, we must choose an appropriate focussing of this lens, i. e. obtain its image by considering that its focus lies at M' and not at F'. Convergent lenses have a negative longitudinal spherical aberration, i. e. nonparaxial rays intersect the axis at a point closer to the lens than the paraxial focus. Divergent lenses have a spherical aberration of the opposite sign.

Spherical aberration can be practically eliminated by a proper choice of the surfaces and systems of lenses. The same is true for the spherical aberration of mirrors.

COMA. If a bright spot forming a wide beam is situated not on the optical axis of the system, its image is not a bright circle as in the previous case, but assumes the form of a quite complicated asymmetric figure. Sometimes, this figure is reminiscent of a comet with a tail, whence the name of this aberration. Coma can be considerably weakened by an appropriate choice of system characteristics.

ABERRATION CAUSED BY OFF-AXIS INCLINED RAYS. The plane passing through the axis of an optical system is called its meridional plane. If a cylindrical beam of rays is incident in this plane at a quite large angle to the axis, it no longer remains cylindrical after refraction. The rays lying in the meridional plane are refracted in a way that is different from the refraction of the rays that are parallel to them but lie off the meridional plane. Consequently, the rays of the beam are not parallel after refraction. Hence the beam cross section varies with distance from the lens after refraction. At a certain distance from the lens, the cross section is a segment of the line perpendicular to the meridional plane (Fig. 80). After this, this line is transformed into an ellipse whose parameters change with distance from the lens. At a certain distance from the lens, the cross section becomes circular, then again elliptical, and finally becomes a segment of a line lying in the meridional plane. Such an aberration is called the astigmatism of inclined beams.

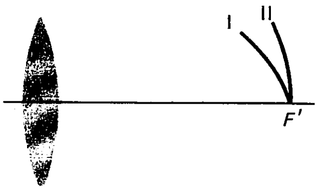


Fig. 83. Meridional and sagittal focal surfaces.

Let us interpret the above-mentioned pattern of refraction of beams in other terms. After passing through a lens, a beam is focussed in the meridional plane and in a plane perpendicular to the meridional plane and parallel to the axis or, in other words, the sagittal plane. The foci for meridional and sagittal focussing are different. In Fig. 80, the meridional focus lies in plane I and the sagittal focus, in plane III. In plane II, the rays from the upper half of the initial cylindrical beam lie in the lower half of the circle, while those belonging to the lower half of the beam lie in the upper half of the circle. The rays from the right half of the initial cylindrical beam lie on the right half of the circle in plane II, while the rays from the left half of the beam lie on the left half of the circle. The position of the planes in which meridional and sagittal focussing takes place depends on the angle of inclination of the incident beam to the optical axis. Hence the surfaces containing the foci formed by meridional and sagittal focussing do not coincide and are not planes. Obviously, these surfaces meet only at the point F' on the optical axis, being mutually orthogonal at this point (Fig. 83). This type of aberration is called the curvature of the image surface. This aberration is removed when Petzval condition is satisfied. We shall not derive or discuss Petzval condition in this book.

The magnification of a system generally depends on the angle of inclination of the ray to the optical axis. For large angles, this dependence becomes significant and the image loses its similarity to the object. As a result, a mesh formed by straight lines is transformed into a mesh of curved lines (Fig. 84). This type of aberration is called distortion.

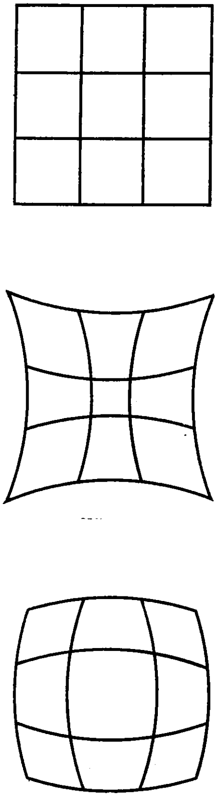


Fig. 84. Distortion.

Geometric aberration can be reduced by an appropriate choice of lenses, their characteristics, and so on. At present, it is possible to remove all aberrations or at least to bring them down to acceptable levels.

CHROMATIC ABERRATION. This aberration is removed by selecting a combination of lenses which would minimize the noncoincidence of images for different wavelengths. It is impossible to obtain an exact coincidence of images for all wavelengths of the spectrum. Usually, an exact coincidence of images is attained for two different wavelengths, and the coincidence for the remaining wavelengths is attained to a certain degree of accuracy. This process is called the achromatization of an optical system. The images for two different wavelengths coincide if the system has the same cardinal elements for these wavelengths. This is attained for the same value of three Gauss constants. In other words, we must have at least three independent parameters in order to attain achromatization. The values of these parameters can be determined as the solution of a system of three equations expressing the condition of coincidence of cardinal elements for both wavelengths. It is always possible to choose an optical system which has three required independent parameters. The problem is simplified by the fact that it is sufficient to achieve a partial achromatization for practical purposes, either by making the foci coincide without bothering about the arrangement of principal planes, or by making the principal planes coincide without bothering about the position of foci. The choice of the achromatization parameters depends on the intended use of the instrument.

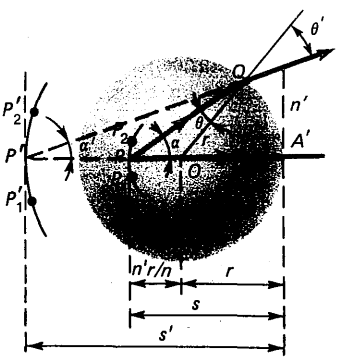


Fig. 85. Image formation for a point lying inside a spherical lens.

Obviously, achromatization can in principle be attained for three or more wavelengths also. For this purpose, we must create a system having a sufficiently large number of free parameters, and choose these parameters appropriately. Achromatization involving more than two wavelengths is also used in optics. The choice of wavelengths for which an exact coincidence of images is attained depends on the purpose for which the instrument is used. For instruments of visual observation, wavelengths close to the maximum sensitivity of the eye, i. e. on either side of the yellow-green spectral region, are used. Usually rays with $\lambda = 656.3$ nm and $\lambda = 486.1$ nm are chosen. For photographic cameras, wavelengths close to the blue region of the spectrum are chosen since the photographic film is sensitive to such wavelengths.

OIL IMMERSION LENS. Oil immersion lenses are employed in order to avoid spherical aberration while using wide beams. The principle of action of such lenses is manifested most clearly in the formation of images of points lying inside a spherical lens (Fig. 85). Point P is chosen at a distance $n'r/n$ from centre O of a spherical lens, where n and n' are refractive indices of the lens and the medium with respect to vacuum. The image is formed as a result of refraction at the spherical surface on which the wave is incident from the concave side. Hence the radius of curvature of this surface appears with a minus sign in the formula. The quantities measured to the left of point A_1 are also negative.

In accordance with the sine theorem, we can write for triangles $P'QO$ and PQO

$$(s' - r)\sin\alpha' = r\sin\theta\,, \tag{24.9}$$

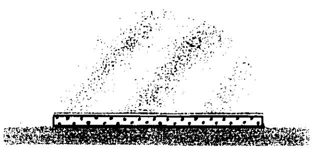


Fig. 86. Construction of an oil immersion lens in a microscope.

$$(s - r)\sin\alpha = r\sin\theta. \tag{24.10}$$

Snell's law of refraction at point Q has the form

$$n'\sin\theta' = n\sin\theta. \tag{24.11}$$

Considering that $s - r = n'r/n$ according to stipulation, we obtain from (24.11) and (24.10)

$$\theta' = \alpha, \tag{24.12}$$

while from triangles $P'OQ$ and POQ having a common angle, we have

$$\alpha + \theta = \alpha' + \theta'. \tag{24.13}$$

Taking (24.12) into account, we find that

$$\theta = \alpha'. \tag{24.14}$$

Consequently, Eq. (24.9) assumes the form

$$s' = r\sin\theta'/\sin\alpha' + r = r\sin\theta'/\sin\theta + r = r(1 + n/n'), \tag{24.15}$$

i.e. s' is independent of angle α. This means that the extensions of all the rays emerging from point P intersect at point P'. In other words, point P' is the image of P. Since the smallness of the angles is not considered in the calculations, there is no spherical aberration in the image formation.

An object cannot be placed inside a glass lens. However, we can create a situation equivalent to placing the object at point P (Fig. 85). For this purpose, the object must be immersed in oil having the same refractive index as the material of the lens, and the front surface of the lens is brought in contact with oil (Fig. 86). This forms the oil immersion lens of a microscope.

It can be seen from (24.14) that the oil immersion lens does not cause spherical aberration and reduces the angles formed by the rays with the axis after leaving the oil immersion lens, since α' is always less than α. This helps in creating for the subsequent elements of the optical system more favourable conditions for the formation of an image without spherical aberration and without any decrease in the intensity of the light flux. By repeated application of this method, we can reduce the angle between the rays and the axis even further. The only restriction on the attainment of indefinitely small angles is the onset of chromatic aberration.

ABBE'S SINE CONDITION. Let us consider the image of segment P_1PP_2 of a circle with radius $n'r/n$ having its centre at point O (see Fig. 85). This image is represented in the form of a segment $P_1'P'P_2'$ of a circle. Moreover, each point of the segment is transformed into a wide beam without spherical aberration. Further, it is necessary that the image of all the points of the segment must have the same magnification. As a result, the image is formed by wide beams without spherical

!

Aberrations are the deviations of the actually obtained images from the ideal ones. The investigation of geometrical aberrations boils down to a consideration of factors which are neglected in paraxial approximation.

Chromatic aberration is caused by the dispersion of light.

aberrations or comas. Let us denote the lengths of arcs P_1PP_2 and $P_1'P'P_2'$ by x and x' respectively. It can be seen from Fig. 85 that

$$\frac{x'}{x} = \frac{s'-r}{s-r} = \frac{s'-r}{rn'/n} = \frac{n \sin \alpha}{n' \sin \alpha'}, \tag{24.16}$$

where we have used equality (24.15). Relation (24.16) in the form

$$x'n' \sin \alpha' = xn \sin \alpha, \tag{24.17}$$

is called Abbe's equation. If an optical system is constructed in such a way that the magnification $x'/x = \text{const}$ for all angles α, the image is formed without spherical aberrations and comas. In the optical system shown in Fig. 85, this condition is satisfied, since on the basis of (24.11), (24.12) and (24.14), we can replace (24.16) by

$$x'/x = n \sin \alpha/(n' \sin \alpha') = n \sin \theta'/(n' \sin \theta) = (n/n')^2 = \text{const}. \tag{24.18}$$

The task of constructing an optical system involves the creation of such conditions that all the rays passing through the system are magnified to the same extent, including the rays which fall from the object on the axis of the system at very large angles. Oil immersion lenses are used to satisfy Abbe's condition for microscopes with a very large magnification.

SEC. 25. OPTICAL INSTRUMENTS

The role of diaphragming is analyzed and the most widespread optical instruments are described.

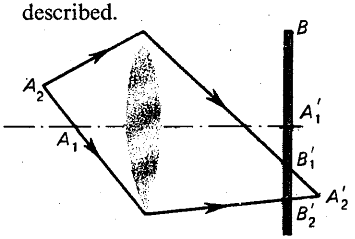


Fig. 87. Illustrating the impossibility of obtaining a plane image of an object with the help of wide beams in an ideal system.

DIAPHRAGMING. To obtain a satisfactory image in optical system, beams of limited width must be used. This is due to the fact that, firstly, the paraxial approximation limits the width of admissible beams. Secondly, even if we assume that an ideal system has been created and the object points are transformed into image points without spherical aberration for unlimited beam widths, the system will not give a satisfactory image of the object unless a restriction is imposed on the beam width (Fig. 87). As a matter of fact, the image must be formed in one plane while the object is three-dimensional and hence ideal images of its points will occupy a certain region. For example, if the image of point A_1 is formed at a certain point A_1' in plane B which is taken as the image plane, the image of point A_2 will be formed at point A_2' which does not lie in the image plane. In the image plane B, the image of point A_2 will be a bright spot whose linear dimension in the plane of the figure is equal to the

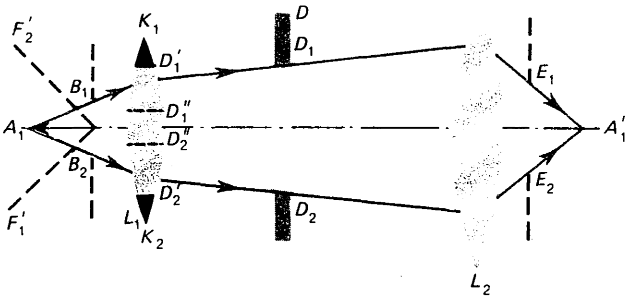


Fig. 88. To the determination of aperture diaphragms of entrance and exit pupils.

length of segment $B_1'B_2'$. The wider the beam of light from point A_2, participating in the formation of the image A_2', the larger the bright spot formed by point A_2 in plane B. For a beam of indefinite width, even an ideal system does not give an image of the object in the image plane. The restriction of the cross sections of light beams is called diaphragming and is attained with the help of diaphragms.

BASIC CONCEPTS ASSOCIATED WITH DIAPHRAGMING. The width of the beams passing through a diaphragmed system is different for different points of the object. For points lying on the axis of an optical system, the diaphragming is characterized by the aperture diaphragm, the entrance pupil and the exit pupil.

The aperture diaphragm is a diaphragm which ensures the maximum restriction of the beam emerging from a point of the object on the optical axis of the system (D_1D_2 in Fig. 88). If lens mount L_1 covered annular regions K_1D_1' and K_2D_2', the aperture diaphragm would be diaphragm D_1D_2 as before. If the lens mount were to cover its annular regions K_1D_1'' and K_2D_2'', the aperture diaphragm would be diaphragm $D_1''D_2''$ and not D_1D_2.

$$M = Dd/(f_1f_2)$$

The entrance pupil is the image of the aperture diaphragm formed by the part of the optical system facing it (B_1B_2 in Fig. 88). If the aperture diaphragm is placed in front of the first lens or is formed by the first lens mount, the entrance pupil coincides with the aperture diaphragm.

The exit pupil is the image of the aperture diaphragm formed by the part of the optical system behind the diaphragm (E_1E_2 in Fig. 88). It can be also stated that the exit pupil is the image of the entrance pupil formed by the entire system. If the aperture diaphragm lies behind the optical system or is formed by the last lens mount, the exit pupil coincides with the aperture diaphragm.

Rays passing through the entrance pupil from off-axis points on the object may be partially or even completely stopped in their path by various parts of the optical system. As a result, the image of these points becomes feeble or even completely disappears. The field of view is the maximum region $F_1'F_2'$ in the object plane, the image of whose points is formed by the optical system without any perceptible image attenuation. The field of view is restricted by a field-of-view diaphragm.

To find the characteristics of the field-of-view diaphragm, the image of each of the diaphragms of the system is formed by the part of the optical system facing the diaphragm. The diaphragm image that can be seen from the centre of the entrance pupil at the largest angle (angle of view) is called the field-of-view diaphragm. Similarly, the field-of-view diaphragm from the image side and the image angle can be determined with respect to the exit pupil.

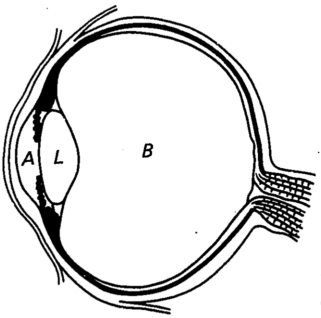


Fig. 89. Eye as an optical system.

EYE AS AN OPTICAL SYSTEM. In the optical sense, the eye is an optical instrument with a variable focal length (Fig. 89). The optical system consists of a set of refractive media including the aqueous humor *A*, crystalline lens *L* and the vitreous body *B*. The change in the focal length occurs through the exertion of muscle force which changes the radius of curvature of the surface of the crystalline lens. The image of the objects is focussed on the retina where the light energy is perceived by the sensory elements of the human nervous system. The information about the distribution of the light flux is transmitted to the brain and appears in the consciousness as the image of the object. It is not expedient to describe the physiological aspects of vision in greater detail in this book.

The focussing of the eye at an object is called accommodation. The most remote point and the nearest point to which the eye can be accommodated are called the far and near points of accommodation. For normal vision, the far point lies at infinity while the near point is at a distance of 10-20 cm from the eye. The range of accommodation decreases with age.

The iris plays the role of the aperture diaphragm. The entrance pupil is the image of the pupil in the anterior part, i. e. in the chamber with aqueous humor. The entrance pupil practically coincides with the pupil of the eye.

As an optical system, the eye can be presented in the form of a reduced eye made of a homogeneous refractive material and having the following characteristics:

Focal power, D	58.48
Length, mm	22
Radius of curvature, mm	
of retina	9.7
of refracting surface	5.6
Refractive index of the medium	1.33

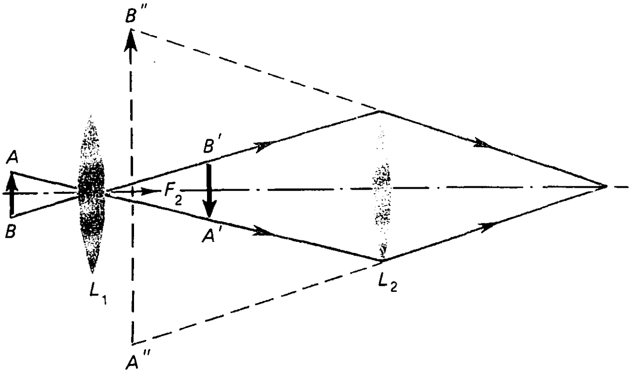


Fig. 90. Microscope.

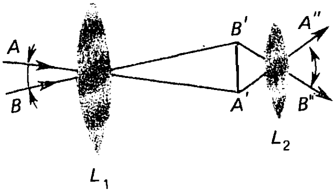


Fig. 91. Telescope.

PHOTOGRAPHIC CAMERA. The optical system of a photographic camera is the objective which ensures a focussing of the image of any object onto a photographic film.

MAGNIFYING GLASS. This is the simplest optical system with a small focal length (1 cm or a bit longer). It consists of one or several lenses. The magnifying glass is placed between the eye and the object. The virtual image is formed either at the near vision point (at about 25 cm), or at infinity. In both cases, the magnification is practically the same and is equal to D/f, where D is the near vision point and f is the focal length of the lens.

MICROSCOPE (Fig. 90). Microscopes are employed for obtaining large magnifications of images of close objects. Real image $A'B'$ of object AB is obtained with the help of objective L_1 and forms a virtual image $A''B''$ in eyepiece L_2. If f_1 and f_2 are the focal lengths of the objective and the eyepiece, and d is the distance between the foci, the magnification of the microscope

$$M = Dd/(f_1 f_2). \tag{25.1}$$

In order to obtain a fairly bright image and to increase the resolving power of the microscope (see Sec. 36), we must use wide beams. Immersion objectives are used in this case to get rid of the spherical aberration (see Sec. 24).

TELESCOPE (Fig. 91). Telescopes are used for obtaining large magnifications of the images of distant objects. The real image of the object formed by long-focus objective L_1 is seen through the eyepiece which serves as the magnifying glass. The magnifying power of the system is

$$M = f_1/f_2, \tag{25.2}$$

where f_1 and f_2 are the focal lengths of the objective and the eyepiece.

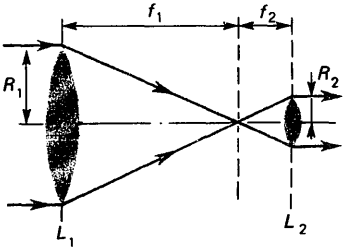


Fig. 92. Path of rays in a telescope.

In telescopes used for observing astronomical bodies, the back focus of the objective can be made to coincide with the front focus of the eyepiece (Fig. 92), as a result of which formula (25.2) can be written as

$$M = f_1/f_2 = R_1/R_2, \tag{25.3}$$

where R_1 and R_2 are the radii of the objective and the eyepiece.

Observations through a telescope are accomplished by the eye. Hence for an optimal utilization of the light flux the exit pupil of the telescope must be of the same size or slightly smaller than the pupil of the eye. For night viewing, the pupil of the eye is about 6-8 mm in size, while for daylight vision its size is just 2-3 mm. If the exit pupil is larger in size than the eye pupil, a part of the light flux is lost on the iris and does not participate in the formation of the image in the eye. Hence for an effective use of the objective surface, the eyepiece should be chosen in such a way that the exit pupil has the optimal size. The magnifying power of the telescope is found to have a fixed value in this case. It is not advisable to use a large objective for a very low magnification. For the same diameter of the objective, it is expedient to use a lower magnification at night than in daylight. Hence if the diameter of the objective is specified, the most effective magnification is found to lie in a very narrow range.

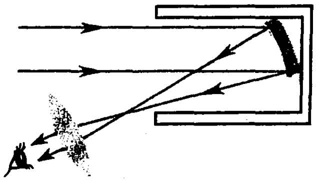


Fig. 93. Construction of a reflecting telescope.

A large angle of view is required in many applications of a telescope (for example, in binoculars). In order to ensure a high quality of the image, we must get rid of the astigmatism of inclined rays, the curvature of the field and chromatism. Hence the eyepieces are usually constructed as complex systems of several lenses.

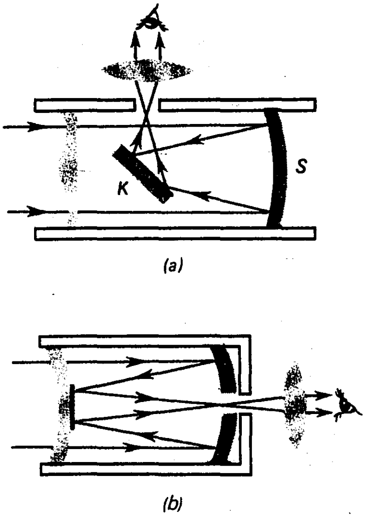


Fig. 94. Schmidt system (*a*) and Maksutov's meniscus system (*b*).

For a fixed size of the exit pupil, the maximum magnification is attained by increasing the diameter of the objective. Moreover, an increase in the diameter of the objective means an enhanced capability of the telescope to distinguish weakly luminous objects. Hence all-out efforts are made to increase the diameters of the telescopic objectives. Making large-diameter lens objectives for telescopes is a very complicated technical problem. It is more convenient to construct a concave mirror of large diameter with a given radius of curvature of its surface. Hence the largest telescopes are not refractors, but reflectors (Fig. 93). Reflectors are free from chromatic aberrations, but suffer from a considerable spherical aberration in case of spherical mirrors. Various methods are employed to eliminate spherical aberration. Nonspherical mirrors, including mirrors whose surface is a paraboloid of revolution, can be used for this purpose. A combination of two nonspherical mirrors gives excellent results.

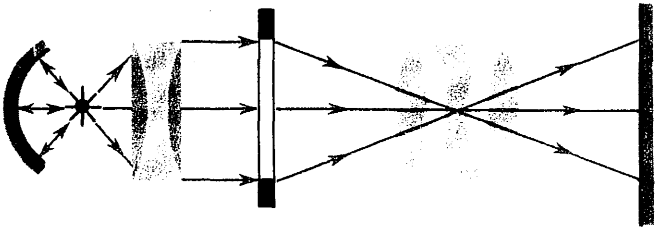


Fig. 95. Optical projector.

It is also possible to use hybrid systems in which lens optics is combined with mirror optics. The longest known system of this type is the Schmidt camera. A correcting plate C is placed in the path of the rays to spherical mirror S (Fig. 94*a*). The plate corrects the phase of the rays incident on it in such a way that spherical aberration and coma are eliminated to a considerable extent. Maksutov's meniscus system is a highly advanced version of this type of telescopes (Fig. 94*b*).

The Maksutov system shown in Fig. 94*b* contains only a spherical mirror and a meniscus, which are bounded by spherical surfaces. This considerably simplifies the calculations and preparation of the system, which is a very important factor in practical applications. At the same time, the system is highly efficient and has high optical parameters.

$M = f_1/f_2$

OPTICAL PROJECTORS. These are intended for producing real magnified images on a screen (slide and film projectors, etc.). The luminous (or illuminated) object to be projected is placed near the principal focal plane of the projector objective which can be moved for sharp focussing (Fig. 95). The object is usually illuminated with the help of a large short-focus condenser having a small focal length so as to admit a considerable light flux through the object being projected. The light flux from the condenser converges at the entrance pupil of the projector objective. The objective and the condenser should be matched in such a way that a sharp image of the object is formed and the light flux passing through the condenser is used to the fullest possible extent.

PROBLEMS

4.1. Find the diameter of the image of the Sun formed by a concave mirror having a radius of curvature of 10 m. The angle at which the diameter of the Sun is seen from the Earth is 32′.

4.2. A coin lies in water at a depth h. At what depth will it seem to be lying to an observer looking at it through water in a vertical direction? The refractive indices of water and air are $n = 1.33$ and $n' = 1$ respectively.

4.3. A microscope with a focal length $f_{ob} = 4$ mm of the objective and $f_{ep} = 25$ mm of the eyepiece is used as a projection device. The distance between the foci of the objective and the eyepiece is $l = 200$ mm. At what distance from the eyepiece must a screen be placed in order to obtain a magnification $M = 2000$ of the object?

4.4. The radius of curvature of the concave surface of a planoconcave lens is 5 cm, and the thickness of the lens is 0.5 cm. A ray is incident on the concave surface of the lens at a distance of 1 cm from its axis, the angle between the ray and the axis being 0.1 rad. Find the distance between the axis and the point at which the ray emerges from the lens, as well as the angle between the emergent ray and the lens axis.

4.5. A double-convex lens (see Fig. 75) with the following parameters is placed in air ($n_1 = 1$): $r_1 = 6$, $r_2 = -3$, $\Delta_1' = 1.5$, $n_1' = 1.5$. Find the focal length of the lens and the position of its cardinal planes.

4.6. An optical system (see Fig. 78) has the following parameters: $f_1 = 30$, $f_2 = -20$, $d = 24$. Find the parameters of the system in the thin-lens approximation. Also determine the magnification.

4.7. A ray emerging from a plane-parallel plate is parallel to the ray incident on the plate. The thickness of the plate and the refractive index of the material are equal to d and n. Find the lateral shift for the ray emerging from the plate.

4.8. A light guide having a circular cross section of diameter d is placed in air. What is the minimum radius of curvature of the outer wall of the light guide? The refractive index of the material of the light guide is n.

4.9. A vertical ray of light is incident on a transparent sphere of radius R and refractive index n at a distance $d < R$ from the vertical axis. For what values of n and d will the point of intersection of the ray and the vertical axis lie inside the sphere?

4.10. A beam of light is incident vertically on a glass hemisphere of radius R lying with its plane side on a table. The axis of the beam coincides with the vertical axis passing through the centre of the base of the hemisphere, and the radius r of the cross section of the beam is smaller than R. Find the radius of the luminous spot formed on the table.

4.11. A rectangular vessel with transparent walls is placed between a double-convex lens and a mirror on which a real image is formed. The vessel is filled with water ($n = 1.33$) and another rectangular vessel with a closed bottom is immersed into the first vessel in such a way that water does not enter the second vessel through its vertical walls. The two opposite plane-parallel walls of the second vessel are perpendicular to the optical axis of the system. Two of the plane-parallel walls of the larger vessel are also perpendicular to the optical axis of the system. A real image of a point lying on the optical axis is then obtained on the mirror. The walls of the vessel immersed into the larger vessel are made of thin glass ($n = 1.5$). By turning the vessel immersed into the larger one about a vertical axis, the angle θ_{in} between the normal to the plane vertical surface of the vessel and the optical axis can be varied. Find the value of the critical angle θ_{cr} at which the image of the point on the mirror disappears.

4.12. A prism is placed in air. The refractive index of the material of the prism is $n = 1.5$, and the base angles are equal to 45°. A ray of light is incident normally on the lateral surface of the prism. Find the transmission coefficient of the prism.

4.13. A circular disc of radius R floats on the surface of water ($n = 1.33$). A small heavy weight is suspended by a very thin thread passing through the centre of the disc and the string is thus stretched vertically downwards. The distance between the water surface and the weight is

h, and the thickness of the disc is negligible. Under what conditions will the weight be visible to an observer in air?

4.14. Find the minimum angle of inclination of a ray emerging from an equilateral prism ($n = 1.5$)? What will the minimum angle of inclination be if the prism is immersed in water ($n = 1.33$)?

4.15. A screen is placed at a distance $l = 2.5$ m from a concave mirror with a radius of curvature $r_2 = -5$ m. At what distance from the mirror must a small object be placed in order to obtain a sharp image on the screen? Find the magnification in this case.

4.16. A narrow beam of light is normally incident on the surface of a transparent sphere in air. For what value of the refractive index of the material of the sphere will the ray be focussed on the opposite surface?

4.17. Derive the formula for the focal length of a system of two thin lenses having focal lengths f_1 and f_2 and separated by a distance d.

4.18. A plane-parallel plate of thickness d is placed between a lens and the image of an object. The refractive index of the material of the plate is $n > 1$. What will the displacement of the image be in this case?

4.19. The objective of a photographic camera is made of two thin lenses, viz. a planoconvex lens made of glass with refractive index n_1, and a double convex lens made of glass with a refractive index n_2. The radius of curvature of the concave surface of the planoconcave lens is R, while the radii of curvature of the surfaces of the double convex lens are R and R'. The surfaces of the lenses having the radius of curvature R are in contact. Find the focal length of the objective.

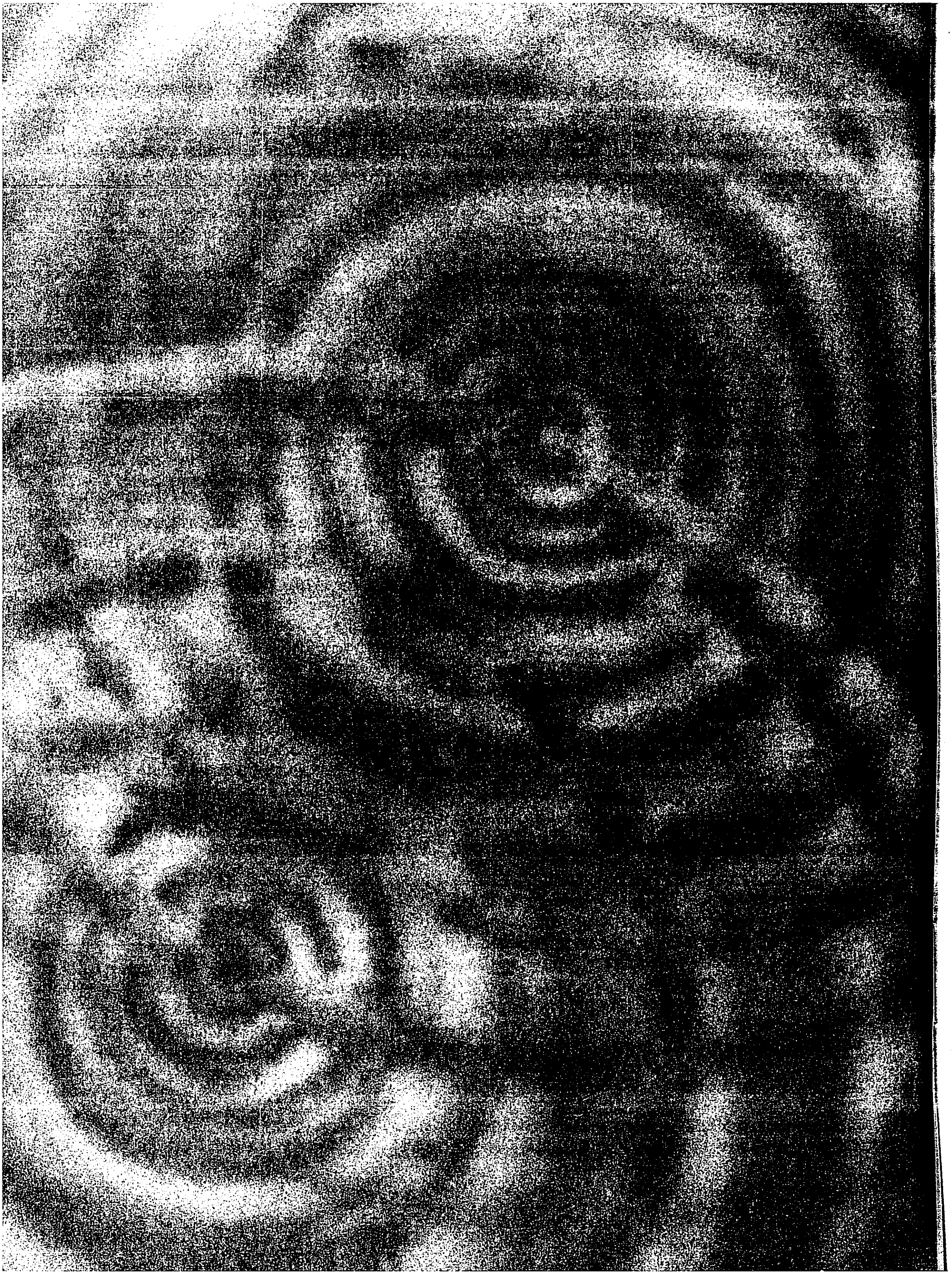
4.20. A telescopic objective consists of two lenses, viz. a convergent lens with a focal length 20 cm and a divergent lens with focal length 5 cm placed behind the first lens at a distance of 15.5 cm. The telescopic objective is directed at a tower of height 30 m at a distance of 3 km. What will the magnification of the image of the tower be and what will the distance between the first lens and the image be?

4.21. A lens is in the form of a spherical ball of radius R. The refractive index of glass is n. Show that the principal planes of this lens coincide and pass through centre of the sphere. Determine focal length of the lens.

4.22. The distance between a concave spherical mirror of radius R and the screen is equal to d. At what distance from the spherical mirror must a small object be placed on the optical axis in order to obtain its sharp image on the screen? What is the magnification in this case?

ANSWERS

4.1. 47 mm. 4.2 $h' = -n'h/n = -0.75h$. 4.3. $Mf_{ob}f_{ep}/l + f_{ep}(1 + f_{ep}/l) = 1.028$ m. 4.4. 1 cm; 0.0 rad. 4.5. $f = 4.26$; $l_H = 0{,}72$; $l'_H = -0.36$. 4.6. $l_H = -51.42$; $l'_H = -34.28$; $l_F = -94.28$; $l'_F = 8.58$; $M = -1.67$. 4.7. $d\sin(\theta_{in} - \theta_{refr})/\cos\theta_{refr}$. 4.8. $d/(n-1)$. 4.9. $(d/R)^2 > 1 - (n^2-2)^2/4$; $\sqrt{2} \leqslant n \leqslant 2$. 4.10. $(r/n)/\{[1-(r/R)^2]^{1/2}[n^2-(r/R)^2]^{1/2} + (r/R)^2\}$. 4.11. 48°36′. 4.12. 0.92. 4.13. $h \geqslant R\sqrt{n^2-1} \approx 0.882R$. 4.14. 37°; 8°27′. 4.15. 0.275 m. $M = -9$. 4.16. 2. 4.17. $1/f = 1/f_1 + 1/f_2 - d/(f_1 f_2)$. 4.18. $-d(1 - 1/n)$. 4.19. $[(n_2 - n_1)/R + (n_2 - 1)/R']^{-1}$. 4.20. 2 cm; 60.5 cm. 4.21. $nR/[2(n-1)]$. 4.22. $dR/(2d - R)$; $-(2d - R)/R$.



Chapter **5**

Interference

Basic idea:

The intensity of a wave resulting from the superposition of two or more waves is determined by the relation between their phases and polarizations.

SEC. 26. TWO-BEAM INTERFERENCE CAUSED BY AMPLITUDE DIVISION

The general formula for two-beam interference is derived and its applications in experiments involving amplitude division are discussed. The visibility of the interference patterns under different experimental conditions is studied. The time coherence is analyzed.

DEFINITION OF INTERFERENCE. Interference is the change in average energy flux density, caused by the superposition of electromagnetic waves. Since [see (3.4)] the energy flux density and the energy density per unit volume are proportional to the square of the amplitude of the electromagnetic wave, and the proportionality factors are constant for a given medium, we shall characterize the energy flux density and the energy density per unit volume of a monochromatic wave by the quantity

$$\boxed{I = \langle \mathrm{Re}E \cdot \mathrm{Re}E \rangle = (1/2)\,\mathrm{Re}(E^* E) = (1/2)\,E_0^2,} \qquad (26.1)$$

called the light intensity. This relation is based on formula (6.19), and E_0 is a real quantity called the amplitude of the light wave. Multiplying (26.1) by ε, we obtain the mean energy density per unit volume, while multiplication by $v\varepsilon$ gives the energy flux density, ε being the permittivity of the medium and v the velocity of light in the medium.

LIGHT INTENSITY UPON SUPERPOSITION OF TWO MONOCHROMATIC WAVES. We shall assume that both waves are linearly polarized in the same direction and have identical amplitudes. They can then be represented in the form

$$E_1 = E_0 \mathrm{e}^{-i(\omega t - \varphi_1)}, \quad E_2 = E_0 \mathrm{e}^{-i(\omega t - \varphi_2)}, \qquad (26.2)$$

where E_0 is assumed to be a real quantity.

As a result of superposition of waves, the electric field strength becomes

$$E = E_1 + E_2, \tag{26.3}$$

while on the basis of (26.1), the light intensity is given by the formula

$$I = \langle \mathrm{Re}(E_1 + E_2)\,\mathrm{Re}(E_1 + E_2)\rangle = (1/2)\,\mathrm{Re}\,[(E_1^* + E_2^*)(E_1 + E_2)] = (1/2)\,\mathrm{Re}(E_1^*E_1 + E_2^*E_2 + E_1^*E_2 + E_2^*E_1), \tag{26.4}$$

where we have used relation (6.19). It follows from (26.2) that

$$E_1^*E_1 = E_0^2,\ E_2^*E_2 = E_0^2,\ E_1^*E_2 = E_0^2 e^{i(\varphi_2-\varphi_1)},$$
$$E_2^*E_1 = E_0^2 e^{-i(\varphi_2-\varphi_1)},$$

and hence

$$I = E_0^2(1 + \cos\delta) = 2I_0(1 + \cos\delta); \tag{26.5}$$

where $\delta = \varphi_2 - \varphi_1$; $I_0 = E_0^2/2$ is the intensity of each of the light waves whose interference is given by formula (26.5).

The relation between the total intensity and the components forming it depends on the phase difference between the waves. If waves of different amplitudes, i.e.

$$E_1 = E_{01}e^{-i(\omega t-\varphi_1)},\quad E_2 = E_{02}e^{-i(\omega t-\varphi_2)}, \tag{26.6}$$

are added, then $E_1^*E_1 = E_{01}^2$, $E_2^*E_2 = E_{02}^2$, $E_1^*E_2 = E_{01}E_{02}e^{i(\varphi_2-\varphi_1)}$, $E_2^*E_1 = E_{02}E_{01}e^{-i(\varphi_2-\varphi_1)}$. Hence formula (26.4) assumes the form

$$I = (1/2)(E_{01}^2 + E_{02}^2 + 2E_{01}E_{02}\cos\delta) = I_1 + I_2 + 2\sqrt{I_1}\sqrt{I_2}\cos\delta, \tag{26.7}$$

where $I_1 = E_{01}^2/2$, $I_2 = E_{02}^2/2$ are the intensities of the composite waves. It can be seen from (26.7) that the total intensity changes from its minimum value

$$I_{\min} = (\sqrt{I_1} - \sqrt{I_2})^2 \quad (\cos\delta = -1) \tag{26.8}$$

to its maximum value

$$I_{\max} = (\sqrt{I_1} + \sqrt{I_2})^2 \quad (\cos\delta = 1). \tag{26.9}$$

METHODS OF PRODUCING COHERENT WAVES IN OPTICS. Two monochromatic waves of identical frequency are required to produce two-beam interference. By definition, such waves have an infinite time duration. It is clear that such waves do not exist in nature. Hence we have to remain content with quasi-monochromatic waves.

$I = (1/2)E_0^2$

Waves suitable for interference can be obtained as a result of division of the same wave into two parts. As regards the time variation of their phase, both waves are exact copies of the original wave. However, a complete analogy with the interference of monochromatic waves is not attained in this case since each wave has a finite coherence time (see Sec. 13) during which these waves can really interfere. Hence the pattern of interference of monochromatic waves is just the first approximation in the investigation of the interference of waves from real sources.

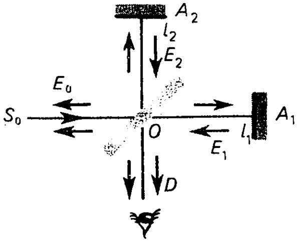


Fig. 96. Amplitude division of a wave in a Michelson interferometer.

There are two ways in which waves can be produced for interference in optics: (1) amplitude division of a wave and (2) division of wave front.

The essence of these methods is shown in Figs. 96 and 97. Figure 96 shows the Michelson interferometer which is used for creating interference by amplitude division of a wave. The wave emerging from source S_0 is incident on a translucent plate O inclined at an angle of 45° to the direction of propagation of the ray. The wave is divided into two parts at the plate: the reflected wave goes towards A_2, while the wave transmitted through the plate goes towards A_1. After reflection from the mirrors A_1 and A_2, the rays are again partially reflected and partially transmitted through plate O. The waves moving towards D can interfere with each other. Obviously, a division of amplitude takes place at plate O, since the wave fronts are conserved at the plate and only the direction of the waves changes.

Figure 97 schematically shows the division of a wave front. The wave from the source is incident on an impenetrable partition containing two holes A_1 and A_2. These parts of the wave front which are incident on the impenetrable part of the partition are reflected or absorbed. Only the parts of the wave front that fall on holes A_1 or A_2 arrive at screen D. Two waves propagate from the holes and can form an interference pattern. Obviously, the screen divides the wave front into parts, and hence the name of this method for producing interference of the waves.

INTERFERENCE OF MONOCHROMATIC WAVES PROPAGATING STRICTLY ALONG THE AXIS OF A MICHELSON INTERFEROMETER. The path difference for the rays is acquired on account of the difference in the lengths of arms l_1 and l_2 of the Michelson interferometer:

$$\Delta = 2(l_1 - l_2). \tag{26.10}$$

The phase difference produced in this case is

$$\delta = 2\pi\Delta/\lambda = k\Delta. \tag{26.11}$$

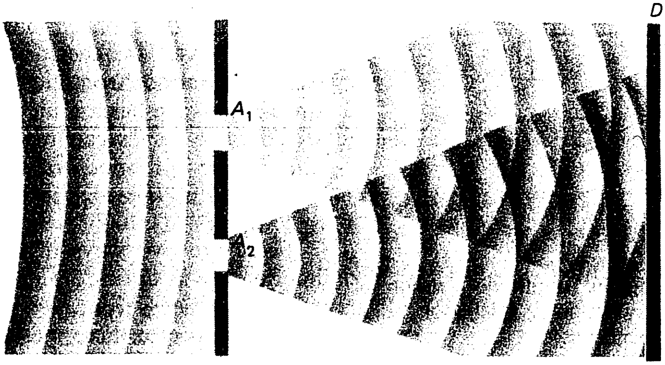


Fig. 97. Division of a wave front at a screen with two holes.

In these calculations, we have not taken into account the phase change of the waves reflected from mirrors *A* and *B*, reflected and refracted by plate *O*, and due to the propagation of light inside the plate. In principle, these factors can be easily calculated. However, we shall not do this since it does not lead to any interesting results as far as the interference is concerned.

We assume that as a result of each incidence of a wave on plate *O*, the energy flux density is divided into two equal parts. If the amplitude of the initial wave is E_0, waves with amplitude $E_0/2$ move in the direction of *D* as a result of two incidences on plate *O*:

$$E_1 = (1/2)E_0 e^{-i(\omega t - \varphi_1)}, \quad E_2 = (1/2)E_0 e^{-i(\omega t - \varphi_2)}, \tag{26.12}$$

Here, $\varphi_2 - \varphi_1 = \delta$. In accordance with (26.5) this gives the following expression for the intensity on the screen *D*:

$$I = (1/4)E_0^2(1 + \cos\delta) = (1/2)I_0(1 + \cos\delta), \tag{26.13}$$

where $I_0 = E_0^2/2$ is the intensity of the wave emitted by the source S_0 and entering the interferometer. Thus, the intensity is equal to zero for $\delta = (2m + 1)\pi\,(m = 0, \pm 1, \pm 2, \ldots)$. This means that there is no energy flux from the plate *O* towards the screen *D*, i.e. the entire energy flux is returned in the direction of the source S_0. If, however, $\delta = 2m\pi\ (m = 0, \pm 1, \pm 2, \ldots)$, the intensity on the screen *D* is equal to I_0, i.e. the entire energy from the source S_0 falls on the screen, and there is no energy flux in the direction of the source S_0. Thus, the energy flux towards *D* depends on δ.

!

The formation of an interference pattern cannot be treated as a result of successive events taking place in the rays.

This can be understood by considering the entire wave field as a whole. The phenomenon described above cannot be considered as a result of successive events taking place in the rays. The situation here is analogous to that in quantum mechanics, which is usually discussed in connection with the

Einstein-Podolskii-Rosen paradox. Hence we shall briefly describe this phenomenon.

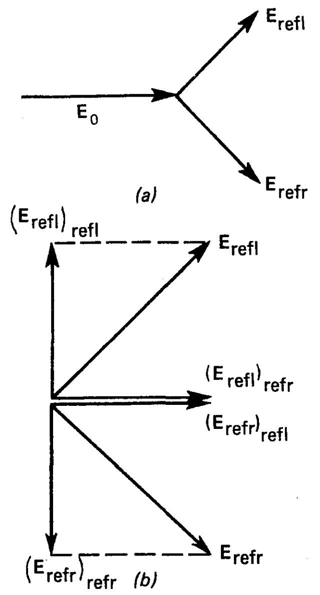


Fig. 98. Vector diagrams of addition of complex amplitudes of rays reflected at a plate and transmitted through it.

Figure 98*a* shows the vector diagram of addition of complex amplitudes of rays reflected at the plate and the rays transmitted through it, taking into account the law of energy conservation for the rays but not the boundary conditions for the waves. Here, E_0, E_{refl} and E_{refr} are the complex amplitudes of the incident, reflected and refracted rays, $|E_{refl}| = |E_{refr}| = |E_0|/\sqrt{2}$, $E_0 = E_{refl} + E_{refr}$. If the path length difference between the rays returning to the plate is an integral multiple of a wavelength, the phase difference between them, shown in Fig. 98*a*, does not change and each of them will be divided into two parts (Fig. 98*b*). The quantities $(E_{refl})_{refl}$ and $(E_{refr})_{refl}$ are the complex amplitudes of reflected rays which were respectively reflected and refracted in the course of their first passage; $(E_{refl})_{refr}$ and $(E_{refr})_{refr}$ are the complex amplitudes of refracted rays which were respectively reflected and refracted upon their first passage through the plate: $E_{refl} = (E_{refl})_{refl} + (E_{refl})_{refr}$, $E_{refr} = (E_{refr})_{refl} + (E_{refr})_{refr}$, $|(E_{refl})_{refr}| = |(E_{refl})_{refl}| = |E_{refl}|/\sqrt{2}$, $|(E_{refr})_{refl}| = |(E_{refr})_{refr}| = |E_{refr}|/\sqrt{2}$.

It can be seen from Fig. 98*b* that $(E_{refl})_{refl} + (E_{refr})_{refr} = 0$, and $(E_{refl})_{refr} + (E_{refr})_{refl} = 2(E_{refl})_{refr} = 2(E_{refr})_{refl}$.

Consequently, there is no ray in the direction towards the source S_0 (see Fig. 96), while in the direction D towards the observer there is a ray with amplitude $|(E_{refl})_{refr} + (E_{refr})_{refl}| = 2|(E_{refl})_{refr}| = \sqrt{2}|E_{refl}| = E_0$, i. e. with the amplitude of the original ray. This corresponds to $\cos\delta = 1$ in Eq. (26.13).

To avoid the concept of wave in this discussion, the concept of phase and amplitude had to be ascribed to a point on the ray. In this case, the division of a ray into two parts contains a contradiction with the phase behaviour upon wave division. The concept of phase and amplitude cannot be referred to a point on the ray. The controversy remains if the ray is divided by taking into account the boundary conditions for the waves. The direction of the complex amplitude E_{refl} in Fig. 98*a* (for the electric field strength perpendicular to the plane of incidence) is then reversed, $E_0 + E_{refl} = E_{refr}$, and the diagram shown in Fig. 98*b* is also changed accordingly. This means that some problems of interpretation associated with quantum mechanics are also inherent in classical physics.

INTERFERENCE OF MONOCHROMATIC WAVES PROPAGATING AT AN ANGLE TO THE INTERFEROMETER AXIS. The easiest way to calculate the path difference between the rays arriving at D and the ones generated by the ray propagating from S_0 at an angle to axis S_0O (see Fig. 96) is to construct the equivalent image of a point source in mirrors

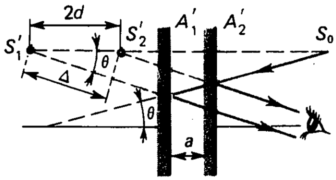


Fig. 99. To the calculation of path difference between rays in a Michelson interferometer.

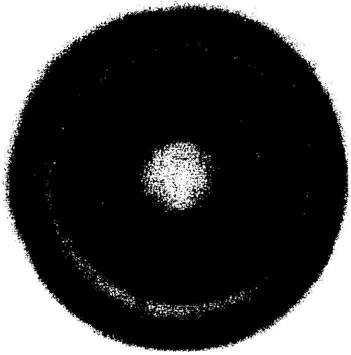


Fig. 100. Variation of the light intensity in a Michelson interferometer.

A_1 and A_2 which are observed at D. All elements of the system are arranged along the same axis for the sake of convenience (Fig. 99). Here, A_1' is mirror A_1 of the Michelson interferometer (see Fig. 96), while A_2' represents mirror A_2 of the interferometer. Point source S_0 (see Fig. 96) is represented by point S_0' in Fig. 99. Points S_2' and S_1' are the images of point S_0 in mirrors A_2 and A_1 respectively. This distance between mirrors A_1 and A_2 is equal to $d = l_1 - l_2$, while the distance between images S_1' and S_2' is twice as much. This gives the following formula for the path difference between the rays:

$$\Delta = 2d\cos\theta = 2(l_1 - l_2)\cos\theta. \tag{26.14}$$

The phase difference between the rays is equal to

$$\delta = k\Delta = (4\pi/\lambda)(l_1 - l_2)\cos\theta. \tag{26.15}$$

The condition of maximum intensity upon interference is given by

$$\delta = 2\pi m \quad (m = 0,\ 1,\ 2,\ 3,\ \ldots) \tag{26.16}$$

or, since $\delta = 2\pi\Delta/\lambda$,

$$\Delta = m\lambda \quad (m = 0,\ 1,\ 2,\ 3,\ \ldots). \tag{26.17}$$

The rays from a point source lying on axis OS_0 diverge at all possible angles θ in an axially symmetric manner. The path difference depends only on angle θ (all the remaining parameters are fixed). Hence the pattern is axially symmetric. The rays emitted at angles corresponding to conditions (26.17) form interference rings of maximum intensity. Thus, the interference pattern consists of alternating bright and dark rings (interference rings, Fig. 100). The integer m in (26.17) is called the order of interference. Substituting the expression for Δ from (26.14) into (26.17), we can write the latter in the form

$$\boxed{2d\cos\theta = m\lambda.} \tag{26.18}$$

It can be seen from here that decreasing values of m correspond to increasing values of angle θ, i.e. rings of large diameters correspond to lower orders of interference. For $m = 1$, Equation (26.18) is satisfied for angles θ close to $\pi/2$. This means that the first order of interference is realized in the form of a ring of very large diameter, which is practically equal to infinity. The observation of interference starts with a large-radius ring, which, however, corresponds to a quite high order of interference. As the order increases, the radius of the interference rings decreases.

As the path difference d increases, the interference fringes become blurred and the interference pattern disappears.

THE REASON BEHIND THE BLURRING OF INTER-

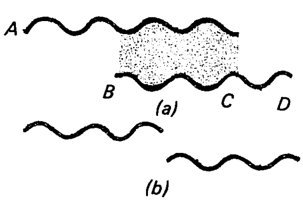
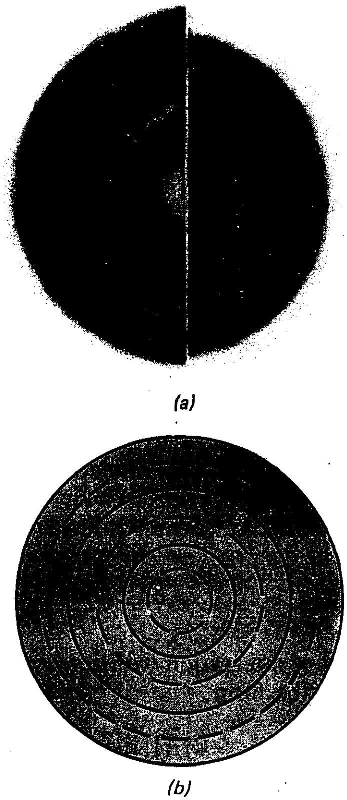


Fig. 101. Interpretation of blurring of the interference pattern due to a wave train of finite length.

FERENCE FRINGES. Upon interference from a strictly monochromatic source of light, no blurring of bands should be observed due to an increase in the path difference. The blurring is caused by finite coherence time and length of the light emitted by the source. This can be explained in two exactly equivalent ways.

If a wave with a finite coherence time and a finite coherence length is interpreted as a harmonic wave of a definite frequency and a finite duration, two interfering waves with a path difference are represented analogously by sinusoidal segments displaced relative to each other in space and time (Fig. 101). The interference contribution, proportional to cos δ, emerges only due to part *BC* of the wave (Fig. 101*a*), while for the remaining time corresponding to segments *AB* and *CD* the interference term is missing and the intensities of the corresponding waves are simply added. Thus, on account of the shift in the sinusoids, the intensity of the interference pattern diminishes and the resultant intensity approaches the sum of the intensities of the two interfering waves. When the path difference becomes so large that the parts of the sinusoids arrive at different times on the screen and do not overlap (Fig. 101*b*), the interference pattern completely disappears.


(a)

(b)

Fig. 102. Interpretation of blurring of the interference pattern due to nonmonochromatic radiation.

Another interpretation of the wave with a finite coherence time and coherence length lies in that the wave is nonmonochromatic and that its spectral width is connected with the time duration of the wave through relation (12.13) in which Δt is now the coherence time. Interference rings of different diameters are formed in the Michelson interferometer due to different frequencies of a nonmonochromatic wave (Fig. 102). Figure 102*a* shows the interference rings of the same order from two waves. The dark rings correspond to interference of waves with shorter wavelengths. It can be seen that the maxima and minima of the interference pattern of waves of different wavelengths are displaced relative to one another, and hence the interference pattern is blurred. If the difference between the wavelengths is sufficiently large, the maxima of the interference pattern of one wave coincide with the minima of the other and the interference pattern completely vanishes (Fig. 102*b*).

Let us quantitatively verify the similarity of these approaches. In the direction given by the angle θ, the interference pattern is completely blurred if the path difference for the incident rays is equal to the coherence length, i.e.

$$2d\cos\theta \propto l_c = c\tau_c .$$

On the other hand, the pattern is blurred if the maxima of the waves with wavelength $\lambda + \Delta\lambda$ coincide with the minima of the waves with wavelength λ. This occurs when

$$2d\cos\theta = m(\lambda + \Delta\lambda) = (m + 1/2)\lambda, \text{ i.e.}$$

$$m\Delta\lambda = \lambda/2,\ 2d\cos\theta = \lambda(\lambda + \Delta\lambda)/(2\Delta\lambda)$$

$$\approx \lambda^2/(2\Delta\lambda) = \lambda\nu/(2\,|\Delta\nu|) \propto c\tau_c,$$

where we have considered that $\Delta\lambda/\lambda = -\Delta\nu/\nu$, and that the relation $\Delta\nu\tau \simeq 1$ is correct only up to a multiplier of the order of unity.

INTERFERENCE OF NONMONOCHROMATIC LIGHT. If the radiation is nonmonochromatic and has a continuous spectrum, the light intensity is proportional to $\mathrm{d}k$ in the interval of wave numbers from k to $k + \mathrm{d}k$:

$$\mathrm{d}I_0 = (1/\pi)F(k)\,\mathrm{d}k. \tag{26.18'}$$

where the factor $1/\pi$ has been introduced in order to match $F(k)$ with the definitions used in the theory of Fourier integrals. In accordance with (26.13), the intensity of light as a result of interference of two waves having a path difference Δ is given by the formula

$$\mathrm{d}I = (1/2)\,\mathrm{d}I_0(1 + \cos\delta) = [1/(2\pi)]\,F(k)(1 + \cos k\Delta)\,\delta k. \tag{26.19}$$

The total intensity is equal to the integral of this expression over all wave numbers:

$$I(\Delta) = \int_{k=0}^{\infty} \mathrm{d}I = [1/(2\pi)]\int_0^{\infty} F(k)(1 + \cos k\Delta)\,\mathrm{d}k. \tag{26.20}$$

The spectrum of waves forming the emission lines is concentrated in a very narrow interval of frequencies in the vicinity of the frequency corresponding to the maximum density of radiation even when the line broadening is relatively small (see Sec. 10). Hence $F(k)$ in (26.20) is nonzero only in a narrow interval of wave numbers near k_0, corresponding to the centre of the emission line. Going over to a new integration variable

$$k' = k - k_0, \tag{26.21}$$

in (26.20), we can change the limits of integration with respect to k' as $\pm\infty$, and write (26.20) in the form

$$I(\Delta) = [1/(2\pi)]\int_{-\infty}^{\infty} F(k_0 + k')\{1 + \cos[(k_0 + k')\Delta]\}\,\mathrm{d}k'$$

$$= [1/(2\pi)]\int_{-\infty}^{\infty} \Phi(k')\{1 + \cos[(k_0 + k')\Delta]\}\,\mathrm{d}k', \tag{26.22}$$

where $\Phi(k') = F(k_0 + k')$. Using the trigonometric relation $\cos[(k_0 + k')\Delta] = \cos(k_0\Delta)\cos(k'\Delta) - \sin(k_0\Delta)\sin(k'\Delta)$, we

!

Interference is determined by the phase difference between the interfering rays.
In optics, coherent rays are obtained by amplitude division and wave front division. An interference pattern can also be formed by two independent laser sources.
The blurring of interference fringes in the Michelson interferometer is due to finite coherence time and length of the light emitted by the source, i.e. to the time coherence.

can transform (26.22) to the equation

$$I(\Delta) = Q + C\cos\beta - S\sin\beta, \tag{26.23}$$

$$\beta = k_0\Delta, \tag{26.24a}$$

$$Q = [1/(2\pi)] \int_{-\infty}^{\infty} \Phi(k')\,\mathrm{d}k', \tag{26.24b}$$

$$C = [1/(2\pi)] \int_{-\infty}^{\infty} \Phi(k')\cos(k'\Delta)\,\mathrm{d}k', \tag{26.24c}$$

$$S = (1/(2\pi)] \int_{-\infty}^{\infty} \Phi(k')\sin(k'\Delta)\,\mathrm{d}k'. \tag{26.24d}$$

It can be seen from (26.23) that light intensity changes harmonically with Δ. It is convenient to express (26.23) in the form

$$I(\Delta) = Q + \sqrt{C^2 + S^2}\cos(\beta + \alpha), \tag{26.25a}$$

where α is defined by the relations

$$\cos\alpha = C/\sqrt{C^2 + S^2}, \quad \sin\alpha = S/\sqrt{C^2 + S^2}. \tag{26.25b}$$

Obviously, the intensity maxima and minima are attained for $\cos(\beta + \alpha) = 1$ and $\cos(\beta + \alpha) = -1$ respectively:

$$I_{\max} = Q + \sqrt{C^2 + S^2}, \quad I_{\min} = Q - \sqrt{C^2 + S^2}. \tag{26.26}$$

The larger the difference between the intensity maxima and minima, the sharper the interference pattern. The sharpness of an interference pattern is quantitatively described by its visibility

$$V = (I_{\max} - I_{\min})/(I_{\max} + I_{\min}). \tag{26.27}$$

The maximum visibility ($V = 1$) is attained for $I_{\min} = 0$, while the minimum visibility is attained for $I_{\max} = I_{\min}$, i.e. when there is no interference pattern. In the general case, we get the following expression for visibility from (26.27) on the basis of (26.26):

$$V = \sqrt{C^2 + S^2}/Q. \tag{26.28}$$

Figure 103*a*, *b* and *c* shows the change in the light intensity I as a function of Δ for different values of the visibility.

FOURIER TRANSFORM SPECTROSCOPY. For $\Delta = 0$, Equation (26.20) assumes the form

$$I(0) = 2[1/(2\pi)] \int_{0}^{\infty} F(k)\,\mathrm{d}k. \tag{26.29}$$

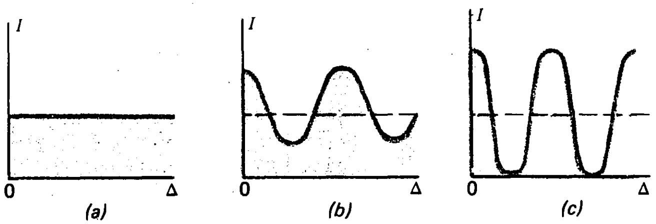


Fig. 103. Dependence of the light intensity on Δ for different values of visibility: $V = 0$ (*a*); $V = 0.5$ (*b*); $V = 1$ (*c*).

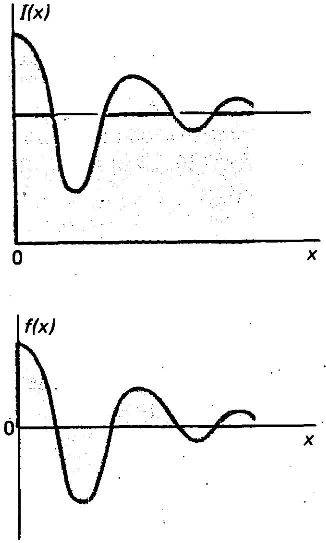


Fig. 104. Relation between the intensity $I(x)$ and the function $f(x)$.

Hence (26.20) can be written as

$$I(x) = 1/2I(0) + [1/(2\pi)] \int_0^\infty F(k) \cos kx \, dk, \tag{26.30}$$

where $x = \Delta$. It follows from (26.30) that the function

$$f(x) = I(x) - I(0)/2, \tag{26.31}$$

called the interferogram, is connected with $F(k)$ through the relation

$$f(x) = \frac{1}{2\pi} \int_0^\infty F(k) \cos kx \, dk, \tag{26.32}$$

which is the Fourier integral for real $F(k)$. The Fourier transform inverse to (26.32) can be written as follows:

$$F(k) = 4 \int_0^\infty f(x) \cos kx \, dx. \tag{26.33}$$

Thus, by presenting the intensity of an interferogram as a function of the path difference $\Delta = x$ of the rays and calculating $f(x)$ from (26.31), we obtain the spectral composition of the radiation with the help of (26.33). This method of spectral analysis is called the Fourier transform spectroscopy (FTS). A variable path difference between the rays is created in a Fourier spectrometer and the corresponding intensity of the interfering rays is measured. The required information on the emission spectrum is then obtained usually with the help of a computer.

It can be seen from (26.31) that curves $I(x)$ and $f(x)$ have identical shape and differ only in the position of the x-axis. By way of an example, Fig. 104 shows the relative positions of these curves. Since the result of experimental measures is $I(x)$ and not $f(x)$, the latter is treated as an auxiliary quantity for spectral analysis, and the spectrum is referred to $I(x)$ and not $f(x)$. The characteristic curves illustrating the relation between

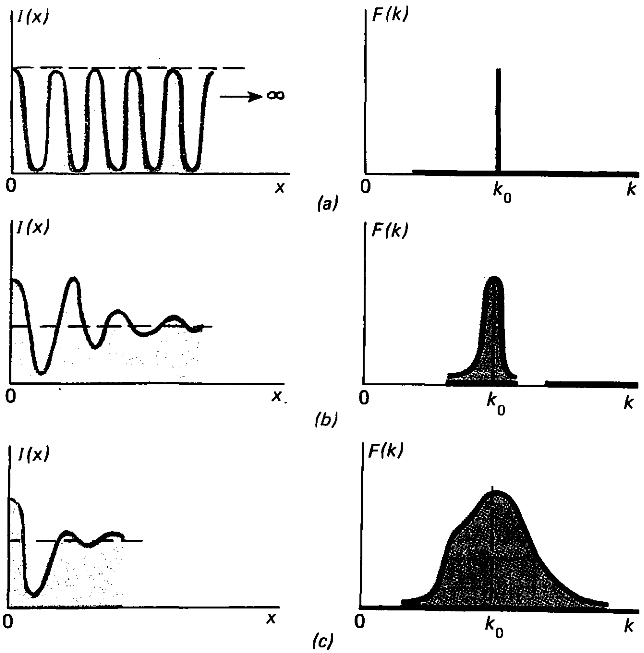


Fig. 105. Relation between the interference signal and the spectrum for monochromatic radiation (*a*); narrow emission line (*b*); broad emission line (*c*).

the spectrum and the intensity of a two-beam interference signal are presented in Fig. 105.

VISIBILITY FOR A GAUSSIAN LINE. Formula (10.17), which describes the Gaussian form of a line as a function of the frequency, can be expressed in terms of the wave numbers ($k = \omega/c$) and presented in the form

$$F(k) = \frac{1}{\sigma'\sqrt{2\pi}} \exp\left[-\frac{(k-k_0)^2}{2\sigma'^2}\right], \tag{26.34}$$

where $\sigma' = \sigma c$ [see (10.17)]. Hence $\Phi(k')$ appearing in formulas (26.22)-(26.24) is given by

$$\Phi(k') = F(k_0 + k') = [1/(\sigma'\sqrt{2\pi})] \exp[-k'^2/(2\sigma'^2)]. \tag{26.35}$$

$$\boxed{V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}}$$

Considering that $\Phi(k') = \Phi(-k')$, we obtain $S = 0$ from (26.24d), and hence

$$V = \frac{C}{S} = \frac{\int\limits_{-\infty}^{\infty} \exp[-k'^2/(2\sigma'^2)] \cos(k'\Delta)\, dk'}{\int\limits_{-\infty}^{\infty} \exp[k'^2/(2\sigma'^2)]\, dk'}$$

$$= \exp(-2\pi^2\Delta^2\sigma'^2), \tag{26.36}$$

which means that the visibility is described by a Gauss function.

VISIBILITY FOR A LORENTZ LINE. In accordance with (9.39), we have

$$\Phi(k') = A/(k'^2 + a^2), \tag{26.37}$$

where $a = \gamma/(2c)$ and A is a constant. Hence $S = 0$ and

$$C = \frac{A}{2\pi}\int\limits_{-\infty}^{\infty}\frac{\cos(\Delta k')\,\mathrm{d}k'}{k'^2 + a^2} = \frac{A}{2a}\mathrm{e}^{-|\Delta a|}. \tag{26.38}$$

Considering that $Q = C(\Delta = 0)$, we obtain the following expression for the visibility:

$$V = C/Q = \exp(-|\Delta a|). \tag{26.39}$$

A comparison of (26.36) and (26.39) reveals that the visibility $V(\Delta)$ of an interference pattern corresponding to a small value of Δ is better for a Gaussian shape of lines, while for large values of Δ the visibility is better for Lorentz forms (Fig. 106).

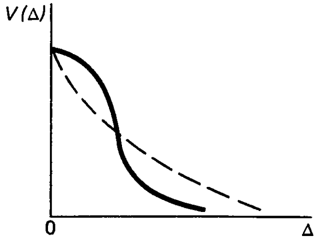


Fig. 106. Dependence of visibility on Δ for radiation with Gaussian (solid curve) and Lorentz (dashed curve) lines.

MICHELSON'S INTERFEROMETER WITH LINEAR FRINGES. It follows from the phase-difference relations

$$\delta = 2\pi\Delta/\lambda = (4\pi/\lambda)(l_1 - l_2)\cos\theta$$

that the distance between the fringes is small for large path differences between the rays, and the fringes merge with one another. A modified version of the Michelson interferometer (Fig. 107) is used for observing interference in this case. The mirror placed in the path of one of the rays in the Michelson interferometer is not strictly perpendicular to the ray. Hence a strictly collimated beam (i.e. a plane wave) is incident on a mirror at a small (grazing) angle, and the path difference depends on the angle between the rays, which varies in the plane of incidence of the ray. Hence the interference fringes are straight lines and do not merge for large path differences.

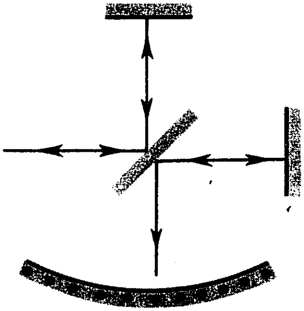


Fig. 107. Modified Michelson interferometer with linear interference fringes.

WHITE LIGHT INTERFERENCE PATTERN. Each wavelength produces its own system of fringes in an interferometer, and the fringes are observed only when the path difference between the rays is small and is just equal to a few wavelengths (small coherence length!). For a zero path difference, the interference from each wavelength takes place with the same magnification and hence a bright spot is observed at the centre. If the path difference $\Delta = \lambda/2$, the waves are suppressed as a result of interference. Hence the bright spot is surrounded, by a dark annular region. An additional path difference emerging as a result of the passage of the rays through the plate dividing them, or by placing additional transparent plates in the path of the rays, may make the central spot dark and the dark annular region bright.

?

What is meant by the visibility of an interference pattern?
What is the underlying principle of the Fourier transform spectroscopy method?
What technique is used for checking the quality of mirrors, lenses and prisms whose surfaces are polished to within a fraction of a wavelength?

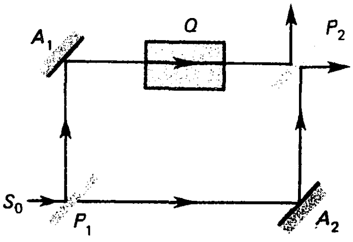


Fig. 108. Mach-Zehnder interferometer.

Beyond the dark annular region, the circular interference fringes formed by different wavelengths begin to overlap, and, as a result, coloured rings are observed. After this the intensity maxima and minima of interference rings of various orders overlap and the whole pattern is blurred. No interference rings are observed in this case.

The order of the observed interference is determined from the interference pattern formed by white light. When monochromatic light is used, the order of interference cannot be found from the interference pattern although the position of the interference fringes can be determined precisely.

In this connection, we shall make certain remarks concerning the adjustment of the Michelson interferometer. In dividing plate O (see Fig. 96) the surface farther from the source is usually the reflecting surface. The distances between the centre of the surface closer to the source and the centres of the mirrors are first made equal with the help of a measuring scale. The ray going towards D after reflection from mirror A_2 passes thrice through the dividing plate, while the ray reflected by mirror A_1 passes through this plate only once. Hence the optical paths are matched by placing in the arm l_1 a transparent plate identical to the dividing plate (but without the reflecting surface) at an angle of 45°. The ray passes twice through this plate. A small ("point") object is placed on the axis of the instrument between source S_0 and O. The two images of this point source formed by mirrors A_1 and A_2 are observed from D. After this, a monochromatic beam of light is directed from the source towards the dividing plate, and the mirrors are turned so that the images of the object coincide, as well as the object with the centre of the emerging system of interference rings. After this, the radius of the interference rings is maximized by adjusting the mirrors. This corresponds to a zero path difference of the rays. Further checking can be made by using white light illumination.

The phenomenon of interference, as manifested by the emergence of colour in thin films, was first discovered by Hooke (1635-1703) and Boyle (1627-1691).

MACH-ZEHNDER INTERFEROMETER. The schematic diagram of this interferometer is shown in Fig. 108. A ray of light S_0 is split into two by a semitransparent plate P_1. After reflection at mirrors A_1 and A_2, the rays of light are recombined by semitransparent plate P_2 as a result of partial reflection and transmission through it. The interference of these rays leads to a pattern similar to the one observed in the Michelson interferometer. If a cell Q containing a gas or a substance with a refractive index other than unity is placed in the path of one of the rays, the interference pattern changes.

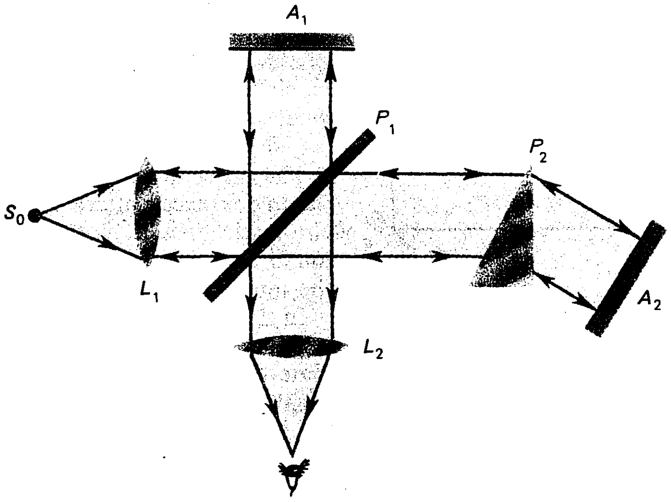


Fig. 109. Twyman-Green interferometer for checking the quality of prism P_2.

The relative refractive index can be measured to a high degree of accuracy in terms of the path of the light ray in the cell and the change in the interference pattern. This makes it possible to study physical processes involving a change in the refractive index.

TWYMAN-GREEN INTERFEROMETER. This is a device used to check the quality of various components of optical instruments. Figure 109 shows a schematic diagram of this interferometer for checking the quality of a prism.

A plane monochromatic wave, formed by monochromatic point source S_0 placed at the focus of lens L_1, is split into two by plate P_1. After reflection at mirrors A_1 and A_2', the two rays are recombined by plate P_1 and are directed towards lens L_2. If plate P_1 and mirrors A_1 and A_2' are optically perfect and maintain a constant path difference between the rays of a beam to within a fraction of a wavelength, the aperture of lens L_2 will be uniformly illuminated. If prism P_2 is optically perfect, the path difference between the interfering rays remains constant over the entire beam cross section, and hence the aperture of lens L_2 will be uniformly illuminated. If, however, the uniformity in illumination is violated and an interference pattern is observed, the prism is not optically perfect. The nature of imperfections is determined from an analysis of the interference pattern and steps are taken to remove these imperfections.

The optical quality of a lens is checked with the help of the Twyman-Green interferometer of the type shown in Fig. 110. The spherical surface of mirror A_2 is polished to within

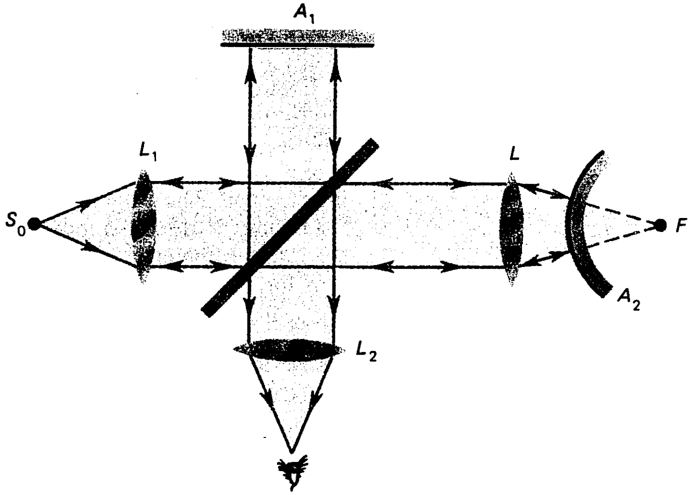


Fig. 110. Twyman-Green interferometer for checking the quality of a lens.

a fraction of a wavelength. The centre of the spherical mirror is made to coincide with focus F of lens L under investigation.

The Twyman-Green interferometer can be used to detect the minutest imperfections in the elements of optical instruments. It is employed for creating nearly perfect mirrors, lenses and prisms.

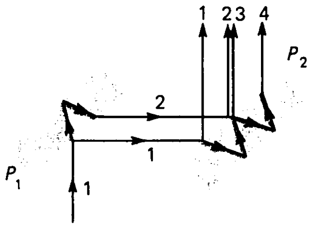


Fig. 111. Jamin refractometer.

JAMIN REFRACTOMETER (Fig. 111). The reflection of rays at the front and back faces of plates P_1 and P_2 results in the formation of four rays *1-4*, two of which (rays *2* and *3*) coincide in space and may interfere. The difference in the paths of the rays reflected at two walls of a thick plate can be determined with the help of Fig. 112. Here, d is the plate thickness and n is the refractive index of the material of the plate relative to vacuum. It can be seen from Fig. 112 that the optical path difference for the rays is

$$\Delta = (|AB| + |BC|)n - |AD|. \tag{26.40}$$

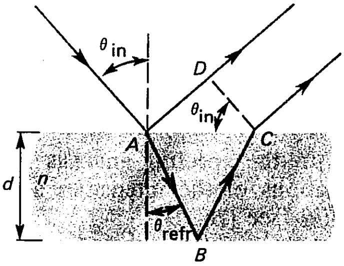


Fig. 112. To the calculation of path difference between rays reflected at the two faces of a plate.

Considering that

$$(|AB| + |BC|) = 2d/\cos\theta_{refr}$$

$$|AD| = 2d\tan\theta_{refr}\sin\theta_{refr},\quad \sin\theta_{refr}/\sin\theta_{in} = n,$$

we get

$$\Delta = 2d/(n/\cos\theta_{refr} - \tan\theta_{refr}\, n\sin\theta_{refr}) = 2nd\cos\theta_{refr}. \tag{26.41}$$

The path difference for the rays in the Jamin refractometer is due to the refraction and reflection of rays in two plates, viz. P_1 and P_2. In accordance with (26.41), this path difference is

$$\Delta = 2nd(\cos\theta_{refr\,1} - \cos\theta_{refr\,2}), \tag{26.42}$$

where $\theta_{\text{refr }1}$ and $\theta_{\text{refr }2}$ are the angles of refraction of the rays in the first and second plates respectively.

Example 26.1. The interference pattern in the Michelson interferometer is observed in the focal plane of a lens with focal length f. Determine radius r_p of a bright ring if the central bright spot is known to correspond to the m_0-th order of interference. The wavelength is denoted by λ.

Obviously, the difference between the lengths of the arms of the interferometer is equal to $l_1 - l_2 = d = m_0\lambda/2$. The condition for the m-th order interference has the form

$$2\mathrm{d}\cos\theta = m\lambda\,. \tag{26.43}$$

Hence it follows that

$$1 - \cos\theta_p = 1 - m/m_0 = (m_0 - m)/m_0 = p/m_0\,, \tag{26.44}$$

where the order of interference decreases with increasing angle θ, and $p = m_0 - m$ is the number of the p-th ring in the interference pattern. Considering that $\theta_p \ll 1$ and taking into account the expansion $\cos\theta_p = 1 - \theta_p^2/2$, we obtain from (26.44) the following expression for angle θ_p of deviation of the ray forming the p-th bright ring:

$$\theta_p = \sqrt{2p/m_0}\,. \tag{26.45}$$

The radius of this ring is given by

$$r_p = 1 - f\theta_p = f\sqrt{2p/m_0}\,. \tag{26.46}$$

If the order m_0 of interference is not known and only the difference in the lengths of the arms in the interferometer is known, we obtain instead of (26.46)

$$r_p = 1 - f\sqrt{p\lambda/\mathrm{d}}\,. \tag{26.47}$$

Example 26.2 The dependence $I(x)$ obtained experimentally by Fourier transform spectroscopy [see (26.30) onwards] can be approximated to a high degree of accuracy by the expression

$$I(x) = A + B\mathrm{e}^{-ax^2}, \tag{26.48}$$

where A, B and a are constants. Find the spectral composition of the radiation.

Since we are interested only in the spectral analysis of radiation, only the relative variations of the intensities are important. Hence we can take $f(x)$ [see (26.31)]:

$$f(x) = (I(x) - B)/A = \mathrm{e}^{-ax^2}. \tag{26.49}$$

Therefore, we obtain in accordance with (26.33)

$$F(k) = 4\int_0^\infty \mathrm{e}^{-ax^2}\cos kx\,\mathrm{d}x = 2\sqrt{\pi/a}\mathrm{e}^{-k^2/(4a^2)} \tag{26.50}$$

i. e. the amplitudes diminish exponentially with increasing frequency. The shape of the spectrum depends only on the relative change in the amplitudes, and hence the absolute value of $F(k)$ in (26.50) is of no significance, the fact which was used in (26.49).

In practical applications of the FTS method, it must be borne in mind that, strictly speaking, (26.50) does not give the exact shape of the spectrum since there are many distortions of the registered intensity (26.48). These distortions are caused by the finiteness of the amplitude, imperfections in the mirrors, noise effects and many other factors. The consideration of all these factors is the main task in the spectral analysis using this method.

SEC. 27. TWO-BEAM INTERFERENCE THROUGH WAVE FRONT SPLITTING

The applications of the general formula for two-beam interference are studied in experiments involving the splitting of the wave front. The role of the source size in interference is discussed and spatial coherence is described.

HUYGENS' PRINCIPLE. While formulating the wave theory of light, Huygens put forth a principle which enabled him to solve some problems associated with propagation and refraction of light in a simple and effective manner. This principle can be stated as follows. If the wave front of a light wave is known at a certain instant of time, the position of the wave front is determined after the interval of time Δt by considering each point of the wave front as a source of a spherical wave and by constructing sphere of radius $c\Delta t$ (c is the velocity of light) about the source. The surface enveloping the secondary spherical waves represents the wave front after an interval Δt of time.

The main drawback of Huygens' principle lies in that it does not take into account the most important property of wave propagation, viz. interference. The principle in which interference is taken into consideration is called the Huygens-Fresnel principle (see Sec. 31).

YOUNG'S TWO-SLIT INTERFEROMETER. The simplest method of splitting a wave front is shown in Fig. 97. In accordance with Huygens' principle, slits A_1 and A_2 can be treated as sources of waves. These sources are generated by the same primary wave and are therefore mutually coherent. Interference is observed between the waves generated by them.

Let us consider the interference produced as a result of separation of two segments of the wave front of the radiation emitted by point source S (Fig. 113) with the help of slits S_1 and S_2. Such an interference set-up was first constructed by Young in 1801.

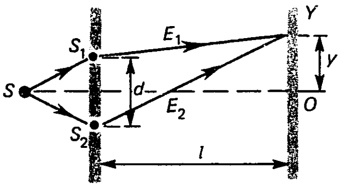


Fig. 113. Schematic diagram of Young's interference experiment.

Figure 113 shows the cross section of the wave front and the

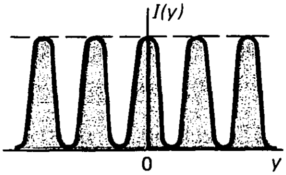


Fig. 114. Variation of intensity upon interference of monochromatic light from two slits (without taking into account the intensity variation due to diffraction).

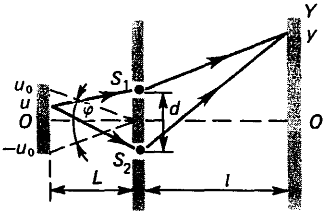


Fig. 115. To the calculation of interference pattern formed by an extended source.

screen with slits by the plane passing through the point source and perpendicular to the screen with slits. For the interference pattern to be identical in all planes parallel to the plane of the figure, we must replace the point source by a linear one in the form of a very long filament perpendicular to the plane of the figure, emitting cylindrical waves. However, the point source will suffice if we are interested only in the interference pattern in the plane of the figure.

While explaining the phenomenon of interference, Young confined himself to qualitative aspect of the problem. Hence his ideas did not receive a wide recognition, and the wave theory of light triumphed only after Fresnel's investigations.

In order to calculate the interference, we must first find the path difference between the rays. It can be seen from Fig. 113 that

$$\Delta = \sqrt{l^2 + (y + d/2)^2} - \sqrt{l^2 + (y - d/2)^2}. \tag{27.1}$$

For $d \ll l$, we can write

$$\sqrt{l^2 + (y \pm d/2)^2} = l[1 + (y \pm d/2)^2/(2l^2)], \tag{27.2}$$

and hence we obtain to within the first-order term in d/l

$$\Delta = dy/l. \tag{27.3}$$

This means that the phase difference between the rays is

$$\delta = 2\pi\Delta/\lambda = [2\pi d/(\lambda l)]y, \tag{27.4}$$

while the change in the intensity with y is given, in accordance with (26.5), by the formula

$$I(y) = 2I_0\{1 + \cos[(2\pi d/(\lambda l))y]\}. \tag{27.5}$$

The plot of variation of the intensity is shown in Fig. 114. It should be noted that this plot does not take into account the diffraction effects at each slit. If diffraction is taken into account, the peak value of the intensity will not be constant (see Sec. 33).

WHITE LIGHT INTERFERENCE. Each wavelength creates its own system of interference fringes, and the central peak is the same for all wavelengths. Hence the central peak is in the form of a bright fringe. The first minima for all the wavelengths are quite close and do not overlap with the higher-order interference fringes. Hence the bright fringe is abutted by dark fringes. For the reasons that were discussed in detail earlier (see Sec. 26), the next interference fringes are coloured. Still further, the fringes are blurred, and the interference pattern disappears.

FINITE-SIZED SOURCE. A source of a finite size can be presented as a combination of mutually incoherent point sources.

Let u be the running coordinate of a point on an extended source relative to axis OO (Fig. 115). For the sake of simplicity, let us consider a linear source. In accordance with (27.3), the path difference between the rays from a point with coordinate u to a point characterized by coordinate y is given by

$$\Delta = (d/l)y + (d/L)u.$$

The phase difference created by this path difference is

$$\delta = (kd/l)y + (kd/L)u \quad (k = 2\pi/\lambda). \tag{27.6}$$

If $i(u)\,\mathrm{d}u$ is the intensity of light emitted by a segment $\mathrm{d}u$ of the source at point u, the contribution from this segment to the overall intensity of the interference pattern at point y is given, in accordance with (27.5), by the formula

$$\mathrm{d}I(y) = 2i(u)(1 + \cos\delta)\,\mathrm{d}u, \tag{27.7}$$

where δ is defined by equality (27.6). This leads to the following expression for the total intensity:

$$I(y) = 2\int_{-\infty}^{\infty} i(u)(1 + \cos\delta)\,\mathrm{d}u. \tag{27.8}$$

Considering that

$$\cos\delta = \cos(k\,dy/l + k\,du/L) = \cos(k\,dy/l)\cos(k\,du/L)$$
$$- \sin(k\,dy/l)\sin(k\,du/L),$$

we can present (27.8) in the form

$$I(y) = 2\int_{-\infty}^{\infty} i(u)\,\mathrm{d}u + 2\cos(k\,dy/l)$$
$$\times \int_{-\infty}^{\infty} i(u)\cos(k\,du/L)\,\mathrm{d}u - 2\sin(k\,dy/l)$$
$$\times \int_{-\infty}^{\infty} i(u)\sin(k\,du/L)\,\mathrm{d}u. \tag{27.9}$$

Using the notation

$$Q = 2\int_{-\infty}^{\infty} i(u)\,\mathrm{d}u,$$
$$C = 2\int_{-\infty}^{\infty} i(u)\cos(k\,du/L)\,\mathrm{d}u, \tag{27.10}$$
$$S = 2\int_{-\infty}^{\infty} i(u)\sin(k\,du/L)\,\mathrm{d}u,$$

we can write this expression in a form similar to (26.23) or (26.25a):

$$I(y) = Q + C\cos(k\,dy/l) - S\sin(k\,dy/l), \tag{27.11}$$

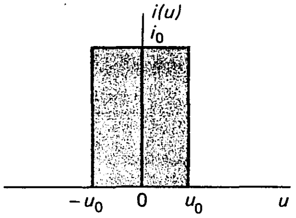


Fig. 116. Brightness distribution in a source of finite size.

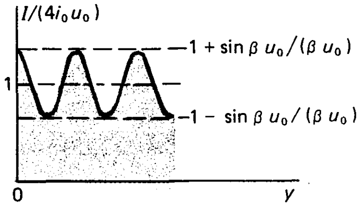


Fig. 117. Intensity distribution in an interference pattern formed by a source of finite size.

$$\underline{I(y) = Q + \sqrt{C^2 + S^2}\cos[(k\,dy/l) + \alpha]}, \tag{27.12}$$

where α is defined by equality (26.25b).

Consequently, the intensity varies according to a harmonic law, and the maxima and minima of intensity are

$$I_{\max} = Q + \sqrt{C^2 + S^2}, \tag{27.13a}$$

$$I_{\min} = Q - \sqrt{C^2 + S^2}. \tag{27.13b}$$

The visibility of an interference pattern is defined by relation (26.27) which can be reduced to the following form with the help of (27.13a) and (27.13b):

$$V = \sqrt{C^2 + S^2}/Q. \tag{27.14}$$

This expression is the same as (26.28).

The analysis of the interference pattern is reduced to the computation of Q, C and S. Let us consider some important cases.

SOURCE WITH A UNIFORM DISTRIBUTION OF EMISSION INTENSITY. The distribution of the brightness is given by the formula

$$i(u) = \begin{cases} 0 & \text{for} \quad -\infty < u < -u_0, \\ i_0 & \text{for} \quad -u_0 < u < u_0, \\ 0 & \text{for} \quad u_0 < u < \infty. \end{cases} \tag{27.15}$$

The plot showing this distribution is presented in Fig. 116. The length of the luminous source is equal to $2u_0$. According to formulas (27.10), we obtain

$$\begin{aligned} Q &= 2\int_{-u_0}^{u_0} i_0\,\mathrm{d}u = 4i_0u_0, \\ C &= 2\int_{-u_0}^{u_0} i_0\cos(k\,du/L)\,\mathrm{d}u = 4i_0u_0\sin(\beta u_0)/(\beta u_0), \\ S &= 2\int_{-u_0}^{u_0} i_0\sin(k\,du/L)\,\mathrm{d}u = 0, \end{aligned} \tag{27.16}$$

where $\beta = kd/L$. As a result, formula (27.11) assumes the form

$$I(y) = 4i_0u_0[1 + \{\sin(\beta u_0)/(\beta u_0)\}\cos(k\,dy/l)]. \tag{27.17}$$

The plot of the intensity distribution is shown in Fig. 117. In accordance with (27.14) and (27.16), the visibility is expressed by the formula

$$V = |\sin(\beta u_0)/(\beta u_0)|. \tag{27.18}$$

For a point source, $u_0 \to 0$ and hence $V \to 1$. According to (27.17), this means that the source described by step function (27.15) produces the same interference fringes as a point source, but with a lower visibility. In other words, the visibility of fringes depends on the size of the source.

This circumstance leads to an important conclusion: the angular dimensions of a source can be determined by measuring the visibility of fringes. This possibility is practically realized with the help of the Michelson stellar interferometer.

SPACE AND TIME COHERENCE. Time coherence must be taken into account while analyzing phenomena in the Michelson interferometer. The analysis of interference caused by the splitting of a wave front requires a consideration of the phase correlation over the wave front at the same intervals of time. This correlation is described by the concept of spatial coherence.

It can be seen from (27.18) that a monochromatic point source ($u_0 \rightarrow 0$) gives an interference pattern with a visibility $V = 1$. A finite-sized source ($2u_0 \neq 0$) consisting of monochromatic point sources that are mutually incoherent gives an interference pattern with a visibility lower than unity. **The radiation emitted by a finite-sized source is monochromatic but not coherent. The degree of coherence of this radiation can be characterized by the visibility of the interference pattern created by this radiation.** If the visibility is equal to zero, the radiation is completely incoherent, if $V = 1$, the radiation is coherent.

For intermediate values of visibility, the radiation is partially coherent.

Coherent radiation (for which $V = 1$) does not exist in actual practice although in some cases (for example, a laser generating a single mode) it is possible to come very close to this limit. Without the help of special radiators, however, this case is just an idealization. Hence we must stipulate the visibility for which the radiation can be called coherent. There is no general consensus regarding this question. According to one opinion, radiation can be called coherent if $V = 0.5$. Other values of V suggested for coherent radiation are 0.9, $\pi/4$, e^{-1}, etc. Depending on the value of V chosen, spatial coherence is characterized by different quantities like the coherence angle and the coherence width. This, however, is not a matter of principle since in different definitions, the values of these parameters differ by multipliers of the order of unity.

COHERENCE ANGLE AND COHERENCE WIDTH. Let us denote by V_c the value of visibility for which the radiation is assumed to be coherent. Hence by taking into account (27.18), we can write the coherence condition in the form

$$|\sin(\beta u_0)/(\beta u_0)| = V_c, \tag{27.19}$$

where $\beta = kd/L$. The solution of Eq. (27.19) in βu_0 can be written in the form

$$\beta u_0 = \theta_0, \tag{27.20}$$

where θ_0 depends on V_c. For example, $\theta_0 = 1.9$ for $V_c = 0.5$.

The coherence angle φ_c is defined as the maximum angular dimensions of a source whose radiation can be assumed to be coherent in accordance with the accepted criterion. Since $u_0 \ll L$ in all cases of interest, the coherence angle can be presented in the form

$$\varphi_c = 2u_0/L, \tag{27.21}$$

where u_0/L must satisfy equality (27.20). Substituting u_0/L from (27.20) into (27.21), we obtain

$$\varphi_c = 2\theta_0/(kd) = (\theta_0/\pi)(\lambda/d). \tag{27.22}$$

For $\theta_0 = 1.9$, we have $\theta_0/\pi = 0.6$. For other values of V_c, we obtain other values of θ_0 and θ_0/π, although all of them have the same order of magnitude. The nature of dependence of the coherence angle on the wavelength and the distance between the centres of the slits remains the same in this case.

Instead of coherence angle, we can use the concept of coherence width, which is defined as the distance along the wave front at which the radiation at the points of the front can be considered as coherent in accordance with the accepted criterion. The distance d_c between the centres of the slits on which the wave front is incident, satisfying this condition, is defined by (27.20). From this equality, we get

$$d_c = (\theta_0 L)/(ku_0) = (\theta_0/2\pi)(L\lambda/u_0). \tag{27.23}$$

For $\theta_0 = 1.9$ we obtain $\theta_0/(2\pi) \simeq 0.3$.

These ideas about spatial coherence, together with those of time coherence described above (see Sec. 26), point towards the complexity of the phenomenon of partial coherence. A suitable description of this phenomenon is possible only in the framework of the general theory of stochastic processes (see Sec. 30). The phenomenon of partial coherence is linked with the theory of stochastic processes due to the physical nature of the radiation. At each point the electric field strength of a wave is a superposition of the electric field strengths of many independent radiators, the frequencies, amplitudes and phases of whose waves are not connected with one another. Hence the total electric field strength is not monochromatic radiation and the change in its amplitudes and phases is of a random nature. Naturally, the magnetic induction of a wave also has similar properties.

The decrease in visibility with an increase in the source size is due to the mutual displacement of interference fringes from different points of the source. As a result, the minimum intensity from some points falls in the region of nonzero light intensity from other points, and the entire interference pattern is blurred. Naturally, the interference pattern will be strongly smeared if the path difference between the rays from the extreme points of the source at a fixed point y on the screen is equal to a half-wave. Hence the coherence condition can be formulated in the form of the requirement that the path difference between the rays from the extreme points of the source be less than a half-wave. This condition can be written in the form

$$\delta\Delta = (d/L)\,\delta u = (d/L)\,2u_0 = \lambda/2,$$

where we have considered that the distance between the extreme points of the source is $\delta u = 2u_0$. Hence the coherence angle is given by

$$\varphi_c = 2u_0/L = \lambda/(2d). \tag{27.24}$$

This leads to the coherence width

$$d_c = \lambda L/(4u_0). \tag{27.25}$$

Formulas (27.24) and (27.25) are in good agreement with formulas (27.22) and (27.23) respectively.

The main physical conclusion from the above description of spatial coherence can be formulated as follows: the radiation from an extended incoherent monochromatic source can be considered as coherent on a surface element whose linear dimensions are of the order of the coherence width and the source of radiation is seen from the centre of this element at the coherence angle. For example, a glowing pin-head held with an extended arm creates coherent radiation on the entrance pupil of the human eye.

STELLAR INTERFEROMETER. MEASUREMENT OF THE DIAMETERS OF STARS. It follows from (27.18) that visibility depends not only on the size $2u_0$ of the source, but also on the distance between the slits. To be precise, it depends on the distance between the slits and the angle at which the source is seen from the midpoint between the slits, since

$$\beta u_0 = k\,du_0/L = (k/2)\,d\varphi, \tag{27.26}$$

where $\varphi = 2u_0/L$ is the angle over which the source is observed. Hence by measuring the visibility of the interference pattern and knowing distance d between the slits, we can determine the angular dimensions of the radiation source, say, a star. This allows us to construct a stellar interferometer for measuring the angular dimensions of astronomical bodies. The idea of a stellar interferometer was first put forth in 1868 by Fizeau.

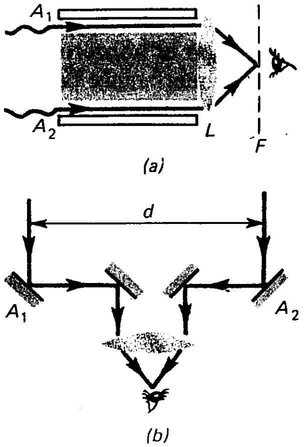


Fig. 118. Rayleigh interferometer (*a*); Michelson stellar interferometer (*b*).

The instrument mainly used for this purpose is called the Rayleigh **interferometer** (Fig. 118*a*). Two parallel beams are separated at slits A_1 and A_2 from a plane wave emitted by a distant star. Upon leaving the instrument, the two parallel beams create in focal plane F of lens L a diffraction pattern which can be used to measure the angular dimensions of the source. However, the Rayleigh interferometer was found to be unsuitable for measuring the angular dimensions of astronomical bodies. The interference fringes obtained in this case are very narrow and it is very difficult to carry out measurements. Fizeau's idea was developed by Michelson who constructed in 1920 an instrument for measuring the angular dimensions of astronomical bodies.

The Michelson **stellar interferometer** is schematically shown in Fig. 118*b*. The slits are replaced by movable mirrors A_1 and A_2 which can be used to change the base d (i.e., the distance between them). The rays reflected by the mirrors are directed after secondary reflection through a lens into an instrument for recording the intensity of interfering rays. The light intensity observed at the centre is defined by formula (27.17) for $y = 0$:

$$I(d) = 4i_0u_0\left[1 + \frac{\sin(k\,d\varphi/2)}{k\,d\varphi/2}\right], \tag{27.27}$$

where we have taken into account expression (27.26). In accordance with (23.18), the visibility of the interference pattern is given by the function

$$V(d) = \left|\frac{\sin(k\,d\varphi/2)}{k\,d\varphi/2}\right|, \tag{27.28}$$

which is plotted in Fig. 119.

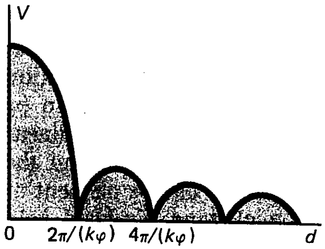


Fig. 119. Visibility of the interference pattern in Michelson stellar interferometer.

As distance d increases from its zero value, the brightness of the central spot gradually decreases. For $d_0 = 2\pi/(k\varphi)$, the spot is completely blurred, and the visibility of the interference pattern becomes equal to zero. From the value of d_0, we can determine the angular dimension of the star under observation:

$$\varphi = 2\pi/(kd_0) = \lambda/d_0. \tag{27.29}$$

Betelgeuse ($\varphi = 0.047''$) was the first star whose angular size was measured in this way. The distance of this star from the Earth was known from terrestrial parallax. The linear size of the star can be calculated from the known values of the distance and angular dimensions. The diameter of Betelgeuse was found to be 300 times more than that of the Sun. The diameters of some other stars and bodies in the solar system were determined in the same way.

!

The visibility of two-beam interference is due to the finite value of the coherence width of a light wave, i.e. to the spatial coherence.
The measurement of the angular dimensions of stars in a stellar interferometer is based on the measurement of the visibility of interference patterns.

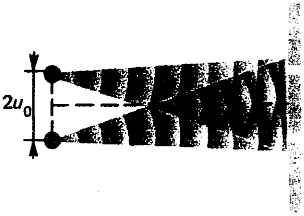


Fig. 120. Interference pattern formed by radiation from a double star.

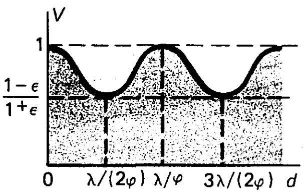


Fig. 121. Visibility of the interference pattern formed by radiation from a double star.

MEASUREMENT OF THE DISTANCE BETWEEN COMPONENTS OF A DOUBLE STAR. If the size of the stars forming a binary system is much smaller than the distance between them, the stars can be treated as point sources and we can consider the interference pattern formed by two point sources separated by distance $2u_0$, i.e. the distance between the double star components (Fig. 120). The brightness of the radiator is distributed according to the formula

$$i(u) = i_0\delta(u + u_0) + \varepsilon i_0\delta(u - u_0), \tag{27.30}$$

where ε characterizes the ratio of the brightness of the components of the system and δ is a delta-function. Taking (27.30) into account, we obtain from formula (27.9)

$$\begin{aligned} I(y) = 2i_0\,[&(1 + \varepsilon) + (1 + \varepsilon)\cos\beta u_0\cos(k\,dy/l) \\ &- (1 - \varepsilon)\sin\beta u_0\sin(k\,dy/l)], \end{aligned} \tag{27.31}$$

where β is the same as in (27.16). Hence we obtain from formula (27.14)

$$V = [(1 + \varepsilon^2) + 2\varepsilon\cos 2\beta u_0]^{1/2}/(1 + \varepsilon). \tag{27.32}$$

By analogy with (27.26), $2\beta u_0 = kd\varphi$, where φ is the angular distance between the components of a double star. The visibility plot $V(d)$ is shown in Fig. 121. As d increases from 0 to $\lambda/(2\varphi)$, the visibility decreases from 1 to $(1 - \varepsilon)/(1 + \varepsilon)$. Our task is to determine the value of d_0 at which the visibility attains its first minimum. The angular distance between the components of a double system is calculated by using the formula

$$\varphi = \lambda/(2d_0). \tag{27.33}$$

FRESNEL BIPRISM. The splitting of a wave front by a Fresnel biprism is shown in Fig. 122*a*. The wave front incident on the biprism is refracted in different directions by the upper and lower prisms. The interference pattern is formed in the region where the refracted wave fronts intersect.

BILLET SPLIT LENS. A lens is cut along its diameter and the two halves are fixed at a certain distance apart. The gap between the separated lenses is covered by an opaque screen. Such a system is called a Billet split lens (Fig. 122*b*). Two real images S_1 and S_2 of a point source S are formed by such a system. The rays diverging from S_1 and S_2 form an interference pattern in the region where they overlap.

LLOYD'S MIRROR INTERFERENCE. Figure 123*a* shows the system of splitting of a wave front and the formation of an interference pattern with the help of Lloyd's mirror. A part of the wave front from source S falls directly on screen

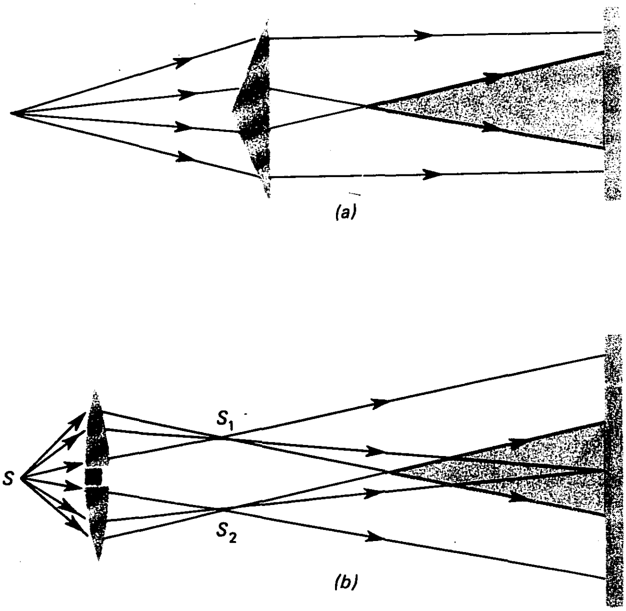


Fig. 122. Fresnel biprism (*a*); Billet split lens (*b*).

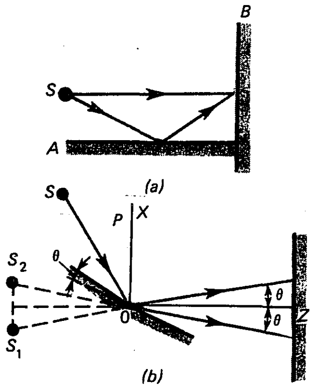


Fig. 123. Lloyd's mirror (*a*); Fresnel mirrors (*b*).

B, while another part is reflected from mirror *A* and is then directed towards the screen. Interference takes place in the region in which the two fronts overlap.

FRESNEL MIRRORS (Fig. 123*b*). Light emitted by a point source *S* is reflected by two plane mirrors inclined at an angle θ, which is close to zero, relative to each other. Interference takes place in the region where the reflected beams overlap. The region of interference between direct rays is covered by a screen *P*. Virtual images of the source are formed at points S_1 and S_2.

LAW OF ENERGY CONSERVATION IN INTERFERENCE. Interference is usually accompanied by a redistribution of energy between various parts of the wave field, the law of conservation of energy being satisfied unconditionally in this redistribution process. For the Michelson interferometer, this aspect was discussed in detail during the analysis of formula (26.13).

This, however, does not mean that the interference itself does not affect the process of emission of electromagnetic waves. It can be seen from the solution of Problem 5.1 that if the distance between two sources of radiation is $\lambda/2$, their total emissive power is not equal to the sum of the individual emissive powers of these sources. Thus, the sources of radiation influence one another, and the emission of radiation is a collective phenomenon. If the distance between the sources is much larger than the length of the waves emitted by them, their mutual interaction is negligibly small.

The total intensity of the interference pattern of a wave field from two sources is generally not equal to the sum of the total intensities of the interference patterns formed by two isolated sources. However, this does not mean that the law of energy conservation is violated. The difference between the field energies is fully explained by the change in the emissive power. For example, if the total intensity increases as a result of interference, the power of the radiators interacting through the field must increase. If the radiator is not in a position to increase the emissive power due to its intrinsic properties, the total intensity of the interference pattern decreases accordingly.

Example 27.1. Analyze the interference produced by the Fresnel mirrors (Fig. 123*b*).

After reflection from the mirrors, the electric field strengths of the waves propagating in the directions indicated by the arrows in Fig. 123*b* are given by

$$E_1 = E_0 \exp[-i(\omega t - kz\cos\theta + kx\sin\theta)], \tag{27.34}$$

$$E_2 = E_0 \exp[-i(\omega t - kz\cos\theta - kx\sin\theta)]. \tag{27.35}$$

Hence

$$E = E_1 + E_2 = E_0 \exp[-i(\omega t - kz\cos\theta)][\exp(ikx\sin\theta) + \exp(-ikx\sin\theta)]. \tag{27.36}$$

Consequently, the interference intensity is given by the formula

$$I = (1/2)EE^* = 2E_0^2\cos^2(kx\sin\theta) = E_0^2[1 + \cos(2kx\sin\theta)]. \tag{27.37}$$

The geometrical meaning of x and $\sin\theta$ is clearly shown in Fig. 123*b*. It can be seen that the interference pattern is exactly the same as if coherent sources (slits) were situated at points S_1 and S_2. Screen P blocks direct rays from source S to the region in which the interference pattern is observed.

?

Under what conditions can the radiation from an extended incoherent source be considered coherent?

What is the reason behind a decrease in the visibility of an interference pattern upon an increase in the size of the source?

SEC. 28. MULTIPLE-BEAM INTERFERENCE THROUGH AMPLITUDE DIVISION

The general method for the analysis of multiple-beam interference is described and its realization is considered for specific interferometer. The resolving power and dispersion region of spectral instruments are defined.

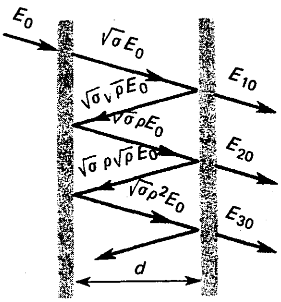

E_0
$\sqrt{\sigma}E_0$
$\sqrt{\sigma}\sqrt{\rho}E_0$
E_{10}
$\sqrt{\sigma}\rho E_0$
$\sqrt{\sigma}\rho\sqrt{\rho}E_0$
E_{20}
$\sqrt{\sigma}\rho^2 E_0$
E_{30}
d

Fig. 124. Path of rays in Fabry-Perot interferometer.

FABRY-PEROT INTERFEROMETER. Let us consider successive partial reflections and refractions of light through two glass plates whose inner surfaces are strictly parallel (Fig. 124), have been polished to a very high precision (from 1/20 to 1/200 of a wavelength), and are covered by highly reflecting films. The films may be metallic (silver, gold, aluminium) or formed by several dielectric layers chosen in such a way as to ensure a very high reflection coefficient (see Sec. 29). The outer surfaces of the glass plates are inclined at a small angle (of the order of 0.1°) to the inner surfaces so that the rays reflected from the outer surfaces are deviated and do not mix with the rays reflected by the inner working surfaces. However, the energy associated with these reflections is insignificant and is not taken into account in subsequent calculations. Moreover, there is no need to take into account the absorption of light as it passes through a glass plate. The **attenuation of the amplitude as a result of reflection is characterized by the reflection coefficient** ρ **[see (18.5)]. The ratio of the amplitudes of the reflected and incident waves is equal to** $\sqrt{\rho}$ (Fig. 124). It is not convenient to use the transmission coefficient τ [see (18.9)] to describe the passage of the wave through the plate since this coefficient connects the amplitude of a wave inside the glass to that of the wave outside the glass, while in the present case it is more convenient to connect the amplitudes of the waves on opposite sides of the glass plate. We denote by $\sqrt{\sigma}$ the ratio of the modulus of the amplitude of the wave transmitted through the plate to the modulus of the amplitude of the incident wave:

$$|E_{0\,\mathrm{tr}}/E_{0\,\mathrm{in}}| = \sqrt{\sigma}. \tag{28.1}$$

In other words, σ is the **transmission coefficient of the plate.** Since the absorption in the plate and the reflection from its outer surface are neglected, the law of conservation of energy is written in a form similar to (8.12):

$$\rho + \sigma = 1. \tag{28.2}$$

On the basis of Fig. 124, we can write

$$\begin{aligned} &E_{10} = \sigma E_0, \ E_{20} = \sigma\rho E_0, \\ &E_{30} = \sigma\rho^2 E_0, \\ &\dots\dots\dots\dots \\ &E_{0N} = \sigma\rho^{N-1}E_0, \end{aligned} \tag{28.3}$$

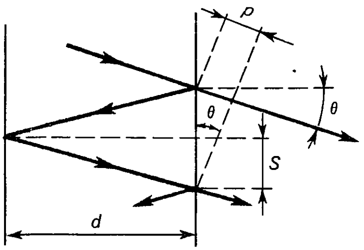


Fig. 125. To the calculation of path difference between neighbouring rays in Fabry-Perot interferometer.

where E_0 is the amplitude of the incident wave, $E_{10}, \ldots, E_{N0}$ are the amplitudes of the mutually interfering waves. The path difference for neighbouring rays can be easily determined with the help of the construction shown in Fig. 125. It can be seen that $\Delta = 2l - p$. Considering that $l = d/\cos\theta$ and $p = 2s\sin\theta = 2(d\tan\theta)\sin\theta$, we obtain

$$\Delta = 2d/\cos\theta - 2d\sin^2\theta/\cos\theta = 2d\cos\theta. \tag{28.4}$$

If the transmitted waves overlap or are condensed to a point with the help of a lens, an interference pattern is observed. The maximum intensity is attained at $\Delta = m\lambda$. The phase difference between neighbouring rays is

$$\delta = k\Delta = (4\pi/\lambda)\,d\cos\theta. \tag{28.5}$$

In accordance with (28.3), the complex amplitude as a result of the interference of N transmitted waves is given by

$$\begin{aligned} E &= \sigma E_0 + \sigma\rho E_0 e^{i\delta} + \sigma\rho^2 E_0 e^{2i\delta} + \ldots + \sigma\rho^{N-1} E_0 e^{i(N-1)\delta} \\ &= \sigma E_0 (1 + \rho e^{i\delta} + \rho^2 e^{2i\delta} + \ldots + \rho^{N-1} e^{(N-1)i\delta}). \end{aligned} \tag{28.6}$$

Since a plane wave is not bounded in space, an infinite number of rays interfere as they emerge out of an interferometer that is not bounded in space. In view of the finite cross section of a light beam and the finite size of the interferometer, however, only a finite number of rays incident at an angle other than zero interfere with one another. The number of interfering waves is always large and so is the number of terms in the geometric progression in (28.6), the last term being much smaller than the first one. Hence we can extend the summation to infinity and write

$$E = \sigma E_0/(1 - \rho e^{i\delta}). \tag{28.7}$$

$$E = \frac{\sigma E_0}{1 - \rho e^{i\delta}}$$

Consequently, the intensity of the interference pattern is

$$I = E^* E/2 = (\sigma^2 E_0^2/2)/(1 + \rho^2 - 2\rho\cos\delta). \tag{28.8}$$

Denoting $E_0^2/2 = I_0$, considering that $1 - \cos\delta = 2\sin^2(\delta/2)$ and taking into account the law of conservation of energy, we can transform (28.8) as follows:

$$I = \frac{I_0}{1 + 4\rho\sin^2(\delta/2)/(1-\rho)^2}. \tag{28.9}$$

INTENSITY DISTRIBUTION IN INTERFERENCE PATTERN. For $\sin(\delta/2) = 0$, i.e. when $\delta = 2\pi n$ ($n = 1, 2, \ldots$), total transmission is observed ($I_{\max} = I_0$) in spite of the fact that each surface through which light passes has a high reflectivity.

The physical reason behind this phenomenon lies in that adjacent waves reflected from each surface have a phase difference equal to π at the first surface and are mutually attenuated. As a result, there is no reflected wave. The phase difference between the adjacent waves at the second surface is 2π; hence they augment each other and pass through this surface without attenuation. For the sake of brevity, we have termed waves with a minimum phase difference as adjacent waves.

For $\sin^2(\delta/2) = 1$, i.e. when $\delta = \pi(2m+1)\,(m = 1, 2, \ldots)$, the intensity of the transmitted light attains its minimum value

$$I_{\min} = I_0(1-\rho)^2/(1+\rho)^2. \tag{28.10}$$

For highly reflecting surfaces, $\rho \simeq 1$ and hence the intensity of the transmitted light is close to zero ($I_{\min} \simeq 0$). This means that practically all the light is reflected. The physical reason behind this phenomenon is that the phase difference between the adjacent waves passing through the second surface is equal to π and they are mutually attenuated. Consequently, there is practically no transmitted wave.

This remarkable phenomenon of almost total transmission or total reflection as an interference effect has an important practical application in the construction of highly efficient mirrors and filters (see Sec. 29).

INTERFERENCE RINGS. If the beam of light incident on an interferometer is formed by rays that are not absolutely parallel, it contains rays with all possible angles θ of incidence. Usually, the angular distribution of rays is axially symmetric. Distance d between the interferometer plates is fixed. Formula (28.5) shows that in this case the phase difference δ is determined only by angle θ (for a given wavelength λ), and hence the interference pattern is axially symmetric. Such a pattern consists of interference rings. Since the rays corresponding to a certain angle θ are parallel to each other, interference rings can be observed on a screen with the help of a lens in its focal plane. In the absence of a screen, the eye is accommodated to infinity.

The interference pattern can be described more conveniently in terms of the phase difference δ rather than the angle θ. These two quantities are connected through relation (28.5). It follows from (28.9) that the visibility of the interference pattern depends on the factor

$$A = 4\rho/(1-\rho)^2. \tag{28.11}$$

This factor depends strongly on reflectivity. For example, $A = 8$ for $\rho = 0.5$ and $A = 360$ for $\rho = 0.9$. Expressing the

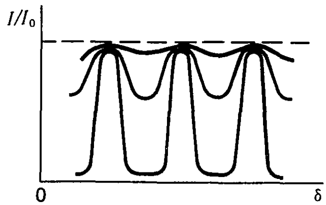


Fig. 126. Variation of intensity in Fabry-Perot interferometer.

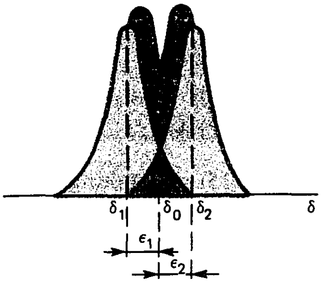


Fig. 127. To the determination of resolving power of optical instruments.

formula (28.9) in the form

$$\frac{I}{I_0} = \frac{1}{1 + A\sin^2(\delta/2)}, \tag{28.12}$$

we conclude that in the first case this ratio varies from 1 to 1/9, i.e. the peak value of the intensity of interference rings is 9 times more than the minimum intensity. In the second case, the intensity varies by a factor of 361. This example illustrates a strong dependence of the visibility of the interference pattern on the reflection properties of the plates. The width of interference peaks also decreases with increasing visibility in accordance with the law of conservation of energy in interference phenomena. The distribution of intensities in an interference pattern for different values of reflectivity, obtained with the help of formula (28.12), is shown in Fig. 126.

RESOLVING POWER. If two waves with wavelengths λ and $\lambda + \Delta\lambda$ are incident on an interferometer, each of them forms its own pattern of interference rings. If the wavelengths are quite close to each other, the two interference patterns nearly coincide and it is impossible to detect the superposition of these two patterns. In this case, it can be stated that the waves with wavelengths $\lambda_1 = \lambda$ and $\lambda_2 = \lambda - \Delta\lambda$ cannot be resolved. The resolving power of an interferometer is characterized by the smallest difference $\Delta\lambda = \lambda_1 - \lambda_2$ between the wavelengths of two waves for which they can be resolved.

Figure 127 shows the distribution of the intensities of waves with close wavelengths for nearly identical values of the path difference:

$$\begin{aligned} \delta_1 &= (2\pi/\lambda_1)\,2d\cos\theta_1 = (2\pi/\lambda)\,2d\cos\theta_1, \\ \delta_2 &= (2\pi/\lambda_2)\,2d\cos\theta_2 = [2\pi/(\lambda - \Delta\lambda)]\,2d\cos\theta_2. \end{aligned} \tag{28.13}$$

These curves have been plotted by using formula (28.12). The values of δ_1 and δ_2 in (28.13) are determined from the condition that they correspond to the peak values of the intensities of the interference rings:

$$\sin(\delta_1/2) = 0, \quad \sin(\delta_2/2) = 0. \tag{28.14}$$

We use the representation of an interference pattern as a function of δ. This is very convenient since it allows a simultaneous analysis of the dependence of the interference pattern on all parameters which determine δ, i.e. the angle θ, the distance d between the plates and the wavelength λ, as well as their combined variation.

Let us establish the conditions under which the intensity distribution of two waves can be assumed as the sum of the intensity distribution of these two waves separately. In the theory of Fabry-Perot interferometer, it is assumed that the resolution condition is the intersection of the intensity distribution curves for each of the waves at half their peak value, i. e. at the point δ_0, where $I/I_0 = 1/2$ for both curves. Hence, in accordance with (28.12), we can write the resolution condition in the form

$$\begin{aligned} &A\sin^2[(\delta_1+\varepsilon_1)/2] = 1\,,\\ &A\sin^2[(\delta_2+\varepsilon_2)/2] = 1\,, \end{aligned} \tag{28.15}$$

where $\varepsilon_1 = \delta_0 - \delta_1$, $\varepsilon_2 = \delta_2 - \delta_0$. It should be observed that ε_1 and ε_2 have very small values since the wavelengths are quite close to each other and the distance $\delta_2 - \delta_1$ between the peaks is very small. We expand the sines in (28.15) into Taylor series in $\varepsilon_1/2$ and $\varepsilon_2/2$ and confine ourselves to the first terms in the expansions:

$$\begin{aligned} &\sin[(\delta_1+\varepsilon_1)/2] = \sin(\delta_1/2) + (\varepsilon_1/2)\cos(\delta_1/2) + \ldots,\\ &\sin[(\delta_2-\varepsilon_2)/2] = \sin(\delta_2/2) - (\varepsilon_2/2)\cos(\delta_2/2) + \ldots. \end{aligned} \tag{28.16}$$

Substituting these expressions into (28.15) and taking into account equalities (28.14) as well as the equalities

$$\cos^2(\delta_1/2) = 1, \quad \cos^2(\delta_2/2) = 1 \tag{28.17}$$

following from them, we obtain instead of (28.15)

$$A(\varepsilon_1/2)^2 = 1, \quad A(\varepsilon_2/2)^2 = 1\,. \tag{28.18}$$

Hence

$$\varepsilon_1 = \varepsilon_2 = 2/\sqrt{A}, \quad \varepsilon_1 + \varepsilon_2 = \delta_2 - \delta_1 = 4/\sqrt{A}\,. \tag{28.19}$$

The angles θ_1 and θ_2 in (28.13), which correspond to the interference rings observed in the Fabry-Perot interferometer, are quite small. Since $\cos\theta = 1 - \theta^2/2 + \ldots$, we can put $\cos\theta_1 = 1$ and $\cos\theta_2 = 1$ in (28.13) with an error not exceeding quantities of the second order of smallness in θ_1 and θ_2. This gives

$$\delta_2 - \delta_1 = 4\pi d\left(\frac{1}{\lambda - \Delta\lambda} - \frac{1}{\lambda}\right) = \frac{4\pi d}{\lambda}\frac{\Delta\lambda}{\lambda}, \tag{28.20}$$

where the approximation $(\lambda - \Delta\lambda)\,\lambda = \lambda^2$ has been used, i. e. while taking into account $\Delta\lambda$ in the numerator, only the zero-order terms in $\Delta\lambda$ have been considered in the denominator. With the help of (28.19), Eq. (28.20) can be transformed as follows:

$$\frac{\Delta\lambda}{\lambda} = \frac{\lambda}{\pi d}\frac{1}{\sqrt{A}} = \frac{\lambda}{2\pi d}\frac{1-\rho}{\sqrt{\rho}}, \tag{28.21}$$

where expression (28.11) has been used for A.

By definition, the resolving power of a spectral instrument is

$$\boxed{\Upsilon = \lambda/\Delta\lambda,} \tag{28.22}$$

where $\Delta\lambda$ is the minimum difference in the wavelengths which can be considered as resolved according to the accepted criteria. On the basis of (28.21), the resolving power of the Fabry-Perot interferometer can be written in the form

$$\Upsilon = \frac{2\pi d}{\lambda}\frac{\sqrt{\rho}}{1-\rho}. \tag{28.23}$$

FACTORS LIMITING THE RESOLVING POWER. By choosing a sufficiently large value of d and the reflectivity ρ close to unity in formula (28.23), we can formally attain an indefinitely large value of the resolving power. However, various circumstances prevent the attainment of indefinitely high resolving power.

In the first place, the interferometer plates cannot be made absolutely flat. It was mentioned above that they can be polished to within 1/200 of a wavelength, but no further. The deviation of the plate surface from an ideal plane is equivalent to a variation of d and leads to a corresponding variation of δ. If d changes by $\alpha\lambda\,(\alpha \ll 1)$, then δ changes by $4\pi\alpha$. To prevent a violation of the resolution condition, this difference must be kept below $\delta_2 - \delta_1$. In other words, the following condition must be satisfied:

$$4\pi\alpha < (\delta_2 - \delta_1) = 2(1-\rho)/\sqrt{\rho}, \tag{28.24}$$

where we have taken into account Eqs. (28.19) and (28.21). Hence very stringent conditions are imposed on the admissible deviation of the plate surface from an ideal plane:

$$\alpha < \frac{1-\rho}{2\pi\sqrt{\rho}}. \tag{28.25}$$

It should be recalled that α characterizes the deviation of the plate surface from an ideal surface in terms of fractions of a wavelength, which indicates that the conditions are indeed very stringent.

The second restriction is imposed by the need to align the plates parallel to each other with the same precision as that achieved in making the plates deviate from an ideal surface. This is obvious from an analysis of the physical meaning of the previous restriction.

The distance between the plates also cannot be made infinitely large. The product of the distance between the plates

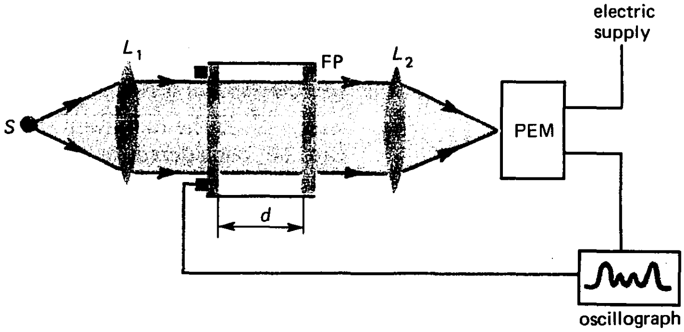


Fig. 128. Fabry-Perot scanning interferometer.

and the number of effective reflections must be less than the time coherence length of the light under investigation.

The attainment of large values of d is also fraught with some practical complications. It follows from (28.5) that the condition for maximum intensity can be presented in the form

$$2d\cos\theta = m\lambda \quad (m = 0,\ 1,\ 2,\ 3,\ \dots). \tag{28.26}$$

This means that the order of interference increases with decreasing diameter of the rings. The central spot ($\theta = 0$) corresponds to the order

$$m_{\max} = 2d/\lambda. \tag{28.27}$$

The first ring has the interference order $m_{\max} - 1$ and is formed by rays whose angle of incidence θ_1 satisfies the condition

$$2d\cos\theta_1 = (m_{\max} - 1)\lambda. \tag{28.28}$$

This is a small angle and, hence, putting

$$\cos\theta_1 = 1 - \theta_1^2/2 \tag{28.29}$$

into (28.28) and taking (28.27) into consideration, we obtain

$$\theta_1 = \sqrt{\lambda/d}. \tag{28.30}$$

This means that as d increases, the radius of the first ring and the distance between the rings decrease and, hence, the analysis of the interference pattern becomes more complicated, especially if photometric technique is employed for this purpose. The analysis becomes much easier if Fabry-Perot scanning interferometer is used (Fig. 128) and if a suitable range of variation of d is chosen.

!

Reflection and refraction depend not only on the properties of the interface and the characteristics of the wave incident on the interface, but also on the characteristics of other waves incident on the interface.

The resolving power of the Fabry-Perot interferometer increases with distance between the plates, while the dispersion region decreases in this case.

DISPERSION REGION. With increasing difference $\Delta\lambda$ between the wavelengths, the interference fringes formed by waves of different wavelengths become separated (Fig. 129*a*). The separation between the fringes increases upon a further increase in $\Delta\lambda$ (Fig. 129*b*), and the interference fringes of adjacent orders overlap for sufficiently large values of $\Delta\lambda$ (Fig. 129*c*).

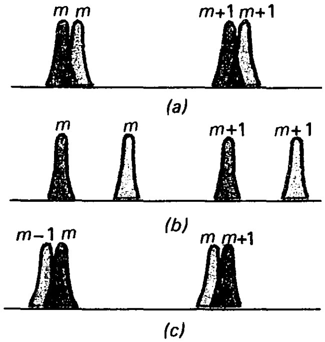


Fig. 129. To the determination of dispersion region.

This complicates the interpretation of the interference pattern. The difference $\Delta\lambda$ between the wavelengths for which the interference fringes of adjacent orders begin to overlap is called the dispersion region. It can be seen from (28.26) that the condition for the coincidence of peaks of adjacent orders can be written in the form

$$2d\cos\theta = (m+1)\lambda, \quad 2d\cos\theta = m(\lambda + \Delta\lambda). \tag{28.31}$$

Hence

$$(m+1)\lambda = m(\lambda + \Delta\lambda) \tag{28.32}$$

or

$$\lambda = m\Delta\lambda, \tag{28.33}$$

and the dispersion region is given by

$$\boxed{G = \Delta\lambda = \lambda/m.} \tag{28.34}$$

For the Fabry-Perot interferometer, $m \approx 2d/\lambda$ for $\theta \simeq 0$, and hence the dispersion region

$$G = \lambda^2/(2d) \tag{28.35}$$

is quite small in view of the fact that higher orders of interference are involved. The requirements of increasing the resolving power and extending the dispersion region are in mutual contradiction: higher interference orders are required for increasing the resolving power (d must be increased), and this naturally reduces the dispersion region.

FABRY-PEROT SCANNING INTERFEROMETER. The measurement of intensity distribution in an interference pattern is extremely laborious. The interference pattern must be photographed and a photometric analysis must then be made in order to obtain the intensity distribution from the darkening of the plate. The temperature fluctuations during the exposure time lead to a displacement of the rings, as a result of which the interference pattern is blurred and the resolution deteriorates. The interference pattern is sensitive to temperature variations of the order of 0.1 K, and a stable temperature must be ensured throughout the experiment.

These difficulties can be overcome by using the Fabry-Perot scanning interferometer (Fig. 128). Light emitted from the point source S placed at the focus of the lens L_1 forms a parallel beam ($\theta \simeq 0$) after passing through the lens. After emerging from the Fabry-Perot interferometer (FP), the beam is focussed on the receiver of the photoelectric multiplier (PEM) where interference takes place. Only the central part of the pattern falls on the receiver of the PEM. The intensity of light incident on the PEM receiver is determined by formula (28.9) in which δ is given by (25.5) with $\cos\theta = 1$. The photomultiplier produces an electric signal whose intensity is proportional to that of light received. This signal is directed to a recording device, say, an oscillograph.

The strength of the signal depends on λ and d. As d changes, the wavelength corresponding to the intensity peak on the PEM receiver also changes. Hence by observing the signal for each value of d, we can directly determine the intensity of the corresponding wave in the radiation incident on the interferometer. One of the plates of the interferometer is mounted on a ring made of a piezoelectric material. The voltage applied to the piezoelectric ring is chosen in such a way that an appropriate variation of d ensures the passage of the entire dispersion region around the wavelength λ corresponding to the peak of intensity at the centre of the interference pattern registered by the PEM receiver. The signal from the photomultiplier is supplied to an oscillograph whose time-base is synchronized with the oscillation frequency of the piezoelectric ring. As a result, the pattern of intensity distribution over wavelengths can be visually observed to a certain scale on the screen of the oscillograph. The observed quantities are then recalculated for the wavelengths and the required emission spectrum is obtained.

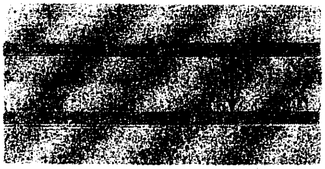


Fig. 130. Interference filter.

INTERFERENCE FILTERS. It was mentioned above that under certain conditions the Fabry-Perot interferometer transmits a light wave of a certain wavelength through it without any attenuation. This is the case when $\sin(\delta/2) = 0$ and $\delta/2 = 2\pi d/\lambda$ $(\cos\theta = 1)$ in formula (28.9). Since the quantity $4\rho/(1-\rho)^2$ in (28.9) is quite large, the denominator becomes very large even for small nonzero values of $\sin(\delta/2)$. Consequently, **the waves whose wavelength differs only slightly from λ are considerably attenuated during their passage (the Fabry-Perot interferometer acts as a narrow-band filter).** In practice, the simplest narrow-band filter is obtained as follows (Fig. 130). A plate made of a dielectric material with a small refractive index is chosen in such a way that the optical path over its thickness d is equal to half the wavelength: $nd = \lambda/2$.

The plate surfaces are covered by thin metallic coatings which have a very high reflectivity but allow a partial transmission of light through them. As a result, we obtain a miniature Fabry-Perot interferometer such that $\delta/2 = \pi$ and $\sin(\delta/2) = 0$ for the chosen wavelength. Hence this system acts as a narrow-band filter for a sufficiently high reflectivity of the metallic coatings. A typical value of the transmission

?

What is the physical reason behind the nearly total reflection or nearly total transmission of a wave in a Fabry-Perot interferometer? Which factors limit the resolving power of the Fabry-Perot interferometer? Why does the Fabry-Perot interferometer have a small dispersion region? What happens to the resolving power of this interferometer when the dispersion region is increased? Describe the principle of action of interference filters.

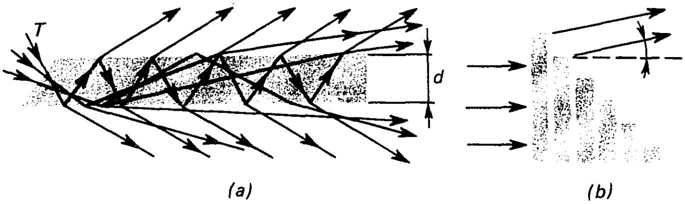


Fig. 131. Lummer-Gehrcke plate (*a*); Michelson echelon grating (*b*).

bandwidth at half the peak intensity is about 20 nm. This is a quite acceptable value for such a simple system. Light is assumed to be normally incident at the surface. The transmission bandwidth can be reduced to half if interference results in second-order maxima. In this case, however, the wave of wavelength 2λ will also pass together with the wave of wavelength λ (see also Sec. 29).

In order to protect the metallic coating from damage, the interference filter is encased in a protective cover. This does not lead to any significant variations in the performance of the filter. In Fig. 130, the protective glass sheets are shown by dots.

LUMMER-GEHRCKE PLATE. The Lummer-Gehrcke plate is a rather simple device for attaining multibeam interference through amplitude division. It consists of a plane-parallel glass plate of thickness d (Fig. 131*a*). Light is made to enter the plate either through a cut at the end face T or through an additional prism, and the entry conditions are chosen in such a way that a ray entering the plate suffers multiple reflection from its faces at angles that are close to the angle of total internal reflection. Denoting by θ the angle of reflection we obtain the following expression for the path difference between the rays as a result of successive reflections from the same face of the plate:

$$\Delta = 2nd\cos\theta. \tag{28.36}$$

This expression is obtained in the same way as (28.4) and is derived by Sec. 29. The condition for the maximum has the form

$$2nd\cos\theta = m\lambda. \tag{28.37}$$

Interference fringes parallel to the plate surfaces are formed in the focal plane of the lens. The plate thickness d usually varies between 3 and 10 mm and the angle θ is close to $\pi/4$. Consequently, the interference order m is found to be very high, i.e. tens of thousands. The interference theory for the Lummer-Gehrcke plate is exactly the same as for the Fabry-Perot interferometer.

MICHELSON ECHELON GRATING. This instrument (Fig. 131*b*) is formed by glass plates of the same thickness, arranged in the form of a staircase with identical steps. The path difference between rays emerging from the relevant points on adjacent steps is due to the difference in the optical path lengths.

Example 28.1. The emission line $\lambda = 546.1$ nm of mercury is a doublet formed by wavelengths λ_1 and λ_2 that are quite close to each other. Assuming that $|1/\lambda_1 - 1/\lambda_2| = 0.1\ \mathrm{cm}^{-1}$, determine the requirements to be met by the reflectivity of the Fabry-Perot interferometer plates and the admissible departures of the plates from an absolutely plane surface. The distance between the plates is 1 mm.

The resolving power of the interferometer is given by formula (28.23). Expressing this quantity in the form $\Delta\lambda/\lambda^2 = \frac{1}{2\pi d}\frac{1-\rho}{\sqrt{\rho}}$ and considering that $|1/\lambda_1 - 1/\lambda_2| = \Delta\lambda/\lambda^2 = 0.1\ \mathrm{cm}^{-1}$, we obtain $\rho = 0.93$. The admissible departure of the plate surface from an ideal plane is given by formula (28.25) and leads to the estimate $\alpha < 1/90$.

SEC. 29. INTERFERENCE IN THIN FILMS

The general methods of multibeam interference are applied for investigating the interference in thin films. The matrix method of analysis and its application to multilayer films, interference filters and dielectric mirrors are described.

OPTICAL PATH LENGTH. The velocity of light in a medium is lower than the velocity of light in vacuum. This means that the wavelength is also lower since the frequency remains unchanged. Consequently, the wave number $k_m = 2\pi/\lambda_m$ in a medium is connected with the wave number k in vacuum through the relation

$$k_m = 2\pi/\lambda_m = 2\pi/(vT) = 2\pi n/(cT) = n2\pi/\lambda = nk, \tag{29.1}$$

where $v = c/n$ is the velocity of the wave in the medium and n is the refractive index of the medium relative to vacuum. The phase difference created due to the passage of the geometrical distance Δ_g by the wave in the medium is

$$\delta = k_m\Delta_g = nk\Delta_g = k\Delta, \tag{29.2}$$

where $\Delta = n\Delta_g$ is the optical path length. Thus, to calculate the phase shift in a medium, it is convenient to assume that the wavelength and the wave vector are the same as in vacuum, while for the path length we take the optical path length which is equal to the product of the geometrical path length and the refractive index of the medium.

REFLECTION FROM PARALLEL SURFACES. Let us denote by n_1 and n_2 the refractive indices of a medium and a film of thickness d (Fig. 132). We shall calculate the difference in the optical paths of the rays reflected from the lower and

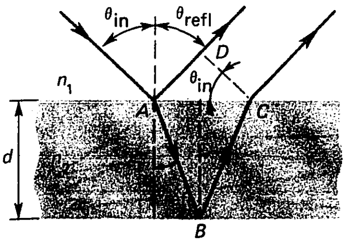


Fig. 132. To the determination of optical path length in thin films.

upper surfaces of the film. It follows from Fig. 132 that

$$\Delta = (|AB| + |BC|)n_2 - |AD|n_1. \tag{29.3}$$

Considering that $|AB| + |BC| = 2d/\cos\theta_{refr}$, $|AD| = 2d\tan\theta_{refr}\sin\theta_{in}$ and taking into account the law of refraction $\sin\theta_{in}/\sin\theta_{refr} = n_2/n_1$, we obtain

$$\Delta = 2d[n_2/\cos\theta_{refr} - \tan\theta_{refr}(n_2/n_1)\sin\theta_{refr}n_1]$$

$$2dn_2\cos\theta_{refr}. \tag{29.4}$$

This formula is identical to (28.4) with the only difference that the geometrical thickness d has been replaced by the "optical thickness" n_2d.

Formula (29.4) shows that all the rays incident on the film at the same angle are divided into two rays which become parallel after reflection from the surfaces of the film and propagate in the same direction. A path difference (29.4) appears between the rays reflected at different surfaces of the film, and hence they can interfere with each other. The interference peak is observed under the condition

$$\Delta = m\lambda. \tag{29.5}$$

It should be emphasized that in this formula λ is the wavelength in vacuum. The peculiarities introduced by the propagation of light in the medium are taken into account by considering that Δ is the optical path length.

Parallel rays do not intersect at finite distances, and hence interference fringes cannot be formed at finite distances. The fringes are said to be localized at infinity. There are two ways in which interference can be observed. The rays are condensed with the help of a lens and the interference fringes can be observed in the focal plane. Alternatively, the fringes can be visually observed by the naked eye accommodated to infinity. In the latter case, the fringes are formed at the focus of the crystalline lens on the retina and create the appropriate visual perception.

The statement that the interference bands are localized at infinity has a mathematical sense as well as a physical meaning. The intensity of the light field in a certain region in space can be measured with the help of a photoelectric cell. By moving the cell in space, we can determine the variation of the intensity of the light field. If the photoelectric cell is shifted in the focal plane of the lens where the interference fringes are formed, it shows the variation in the intensity of the light field in accordance with the observed interference pattern. If, however, the photoelectric cell is shifted in the space while the interference fringes are localized at infinity, no variation of the light field due to interference will be observed.

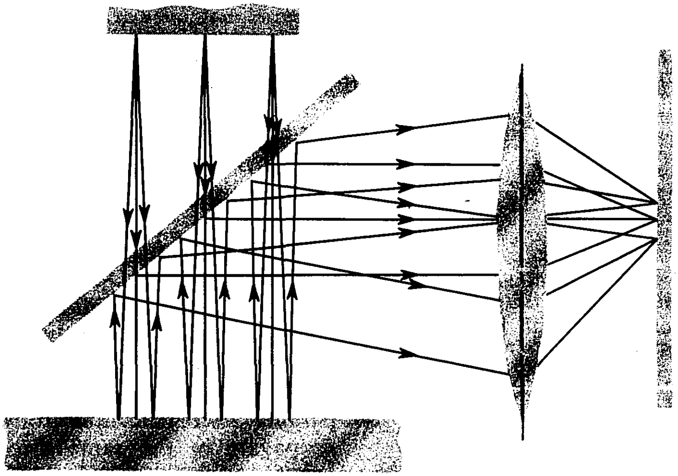


Fig. 133. Formation of uniform inclination interference fringes on a screen.

UNIFORM INCLINATION FRINGES. If a beam of nonparallel rays is incident on a plate, the reflected beam will contain rays propagating in different directions after being refracted at different angles. The rays satisfying relation (29.5) give an intensity maximum as a result of interference. Hence if an interference pattern is formed in the focal plane of a lens, the interference fringe of a certain intensity corresponds to a definite angle θ_{refr} in (29.4), i.e. to a definite angle of incidence or reflection. In other words, this fringe corresponds to a definite inclination of the rays forming it to the plate surface. Hence such interference fringes are called uniform inclination fringes (isoclinic fringes). These fringes are localized at infinity.

If an axisymmetric beam of light is incident on a film, the uniform inclination fringes will be circles. Figure 133 shows how interference fringes of equal inclination are produced on a screen with the help of a lens.

$\Delta = (|AB| + |BC|)n_2 - |AD|n_1$

Uniform inclination fringes which are circles in the case of an axisymmetric incidence on a plate can be seen by the naked eye (Fig. 134). If as a result of the displacement of a film parallel to itself its thickness changes in the visible region, the radius of the interference rings also changes and this change is easy to notice for quite insignificant changes in the film thickness. This provides a very effective method of checking the thickness of a film (plate) for industrial application. The accuracy of this method is very high since local variations in the thickness of a plate by a fraction of a wavelength leads to a significant variation in the interference pattern which can be observed by the naked eye.

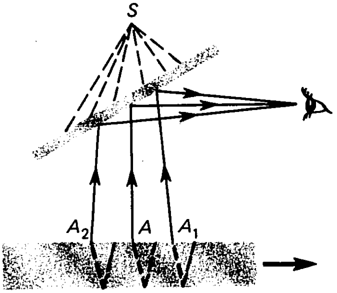


Fig. 134. Observation of uniform inclination fringes by the naked eye. Method of checking plate thickness.

THE ROLE OF SOURCE SIZE. The emission of radiation from different points of a source is incoherent. However, interference patterns formed by any point of the source remain identical upon reflection at the same angle and do not depend on the point of the film where the reflection occurred. The interference patterns formed by the radiation emitted by different points of the source overlap without blurring the interference pattern. Hence the finite size of the source does not blur the interference pattern of uniform inclination fringes and is not a factor which imposes any restrictions on the interference.

THE ROLE OF PLATE THICKNESS AND MONOCHROMATICITY OF RADIATION. As the plate thickness is increased, the visibility of the interference pattern is reduced in view of the finite value of time coherence. The reason behind this is similar to the one considered in the analysis of deterioration of visibility in a Michelson interferometer (see Sec. 26). An increase in the width of the emission line (nonmonochromaticity) also deteriorates the visibility of the interference pattern.

The necessary condition for observing interference in thin films is that their thickness must not exceed a certain value. When interference is observed in white light, the overlapping of fringes from different wavelengths must not smear the colour spectrum of white light. It is well known that the human eye can distinguish between colours corresponding to a difference of about $\Delta\lambda \simeq 10^{-2}$ μm in the wavelengths. Hence a dispersion region of the same size must be ensured if interference pattern is to be observed. It follows from (28.34) that the order of interference must not exceed

$$m_{\max} = \lambda/\Delta\lambda = 0.5 \times 10^{-6}/10^{-8} \simeq 50, \tag{29.6}$$

where $\lambda = 0.5$ μm by way of an example. From condition (29.5) we obtain

$$d \approx m\lambda/(2n_2) = [50 \times 0.5 \times 10^{-6}/(2 \times 1.5)]\ \text{m}$$
$$= 8 \times 10^{-6}\ \text{m} = 8\ \mu\text{m}, \tag{29.7}$$

where $\cos\theta_{\text{refr}} = 1$ and $n = 1.5$. Of course, this calculation can be carried out on the basis of the time coherence. It follows from the relation $\lambda = c/\nu$ that $\Delta\lambda/\lambda = -\Delta\nu/\nu$. The spectral width $\Delta\nu$ is connected with the coherence time τ through the relation $\Delta\nu\tau = 1$. The velocity of propagation of light in the plate is c/n_2 and the distance traversed in time τ must not be less than $2d$. Hence we obtain the relation $2d = c\tau/n_2 = c\lambda/(\Delta\lambda\nu)$ for determining the maximum plate thickness. From this relation we obtain $d = \lambda^2/(2\Delta\lambda n_2) = m\lambda/(2n_2)$, which coincides with (29.7). Thus, the film thickness must be quite small if interference is to be observed in white light. Hence we refer to interference in thin films and not just in plane-parallel plates.

When interference from a monochromatic source is observed, the admissible thickness of the film increases. It depends on the width of the emission line or, in other words, on the coherence time and coherence length. For example, the width of the green emission line from a mercury lamp in terms of wavelength is $\Delta\lambda \simeq 0.01$ nm ($\lambda = 502.564$ nm). Hence, in accordance with formulas (29.6) and (29.7), the maximum admissible film thickness is $d = 8$ mm.

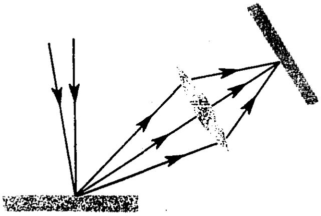


Fig. 135. Observation of uniform thickness fringes.

For laser radiation, the coherence length may attain very large values (several kilometres). Hence it can be used for observing interference patterns in very thick plates. The possibility of observing the patterns is limited not by the degree of coherence but by the inhomogeneities of the material, the quality of the plate surfaces and other related factors.

UNIFORM THICKNESS FRINGES (ISOPACHIC FRINGES). If a plate has a varying thickness, pairs of rays with the same phase difference propagate in different directions from various parts of the plate surface. Hence the interference pattern formed by uniform inclination fringes is not observed. However, another interference pattern which is localized on the surface of the plate can be seen. The interference fringes forming this pattern are called isopachic fringes in view of the fact that the fringes have the same intensity in the regions where the plate thickness is the same.

We focus a lens in such a way that the image of a small part of the surface is obtained on the screen (Fig. 135). The illuminance of the region depends on the plate thickness in this region since rays with almost the same path difference interfere for nearly normal incidence at the plate surface. The path lengths of the rays arriving at a given point on the surface differ only because of their different angles of refraction (θ_{refr}). However, this difference is not significant since the linear size of the region under consideration is quite small. Hence the interference pattern projected by the lens on the screen is the one formed at the plate surface. If a photoelectric cell is moved over the surface of the plate, it will register the variation in the light field intensity corresponding to the change in the intensity of the interference pattern being observed. The bright and dark fringes on the surface of the plate can also be seen directly by the eye accommodated to the plate surface. The fringes of uniform interference intensity coincide with the fringes of equal thickness of the plate (isopachic fringes). The region of uniform intensity corresponds to the region in which the plate thickness is the same. The contrast in the interference pattern depends on the order of interference, i. e. on the plate thickness. Coloured fringes are observed when white light is used for illumination. The limitations on the plate thickness are the same as in the case of uniform inclination fringes just considered.

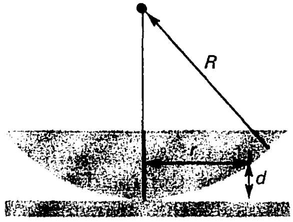


Fig. 136. To the calculation of the radius of Newton's rings.

NEWTON RINGS. Isopachic fringes are formed due to the air space between the plane surface of a glass plate and planoconvex lens which rests on the plate with its convex side downwards (Fig. 136). The isopachic fringes of thickness d in the air space are circles. Hence the interference fringes of uniform thickness are also circles. These fringes are called Newton's rings.

It can be seen from Fig. 136 that $R^2 - (R - d)^2 = r^2$, $2Rd - d^2 = r^2$. Writing this relation in the form $2Rd/r^2 - d^2/r^2 = 1$ and considering that $d^2/r^2 \ll 1$, we obtain

$$d = r^2/(2R). \tag{29.8}$$

Upon reflection at the boundary with a denser medium, the phase of the electric field strength of a wave changes by π [see (16.33a), Fig. 65]. Hence the condition for the formation of a dark ring has the form

$$2r_m^2/(2R) + \lambda/2 = m\lambda + \lambda/2.$$

This gives the following expression for the radius of the m-th dark ring:

$$r_m = \sqrt{R\lambda m}. \tag{29.9}$$

A dark spot is observed in the region where the lens is in contact with the glass surface. This spot is surrounded by alternating bright and dark rings. When white light is used for illumination, coloured rings are observed. In principle, the conditions for observation of the rings are the same as for thin films. In other words, the interference pattern is blurred with increasing d. In transmitted light, an additional interference pattern is observed. Among other things, the bright rings are replaced by dark ones, and vice versa, when monochromatic light is used for illumination.

MULTIPLE REFLECTION. In the analysis of interference in thin films so far, we have only considered the two-beam interference resulting from a single reflection at the film surface (see Fig. 132). Such an approximation gives good results and is thoroughly justified if the reflection coefficient at the film surface is small. When the reflectivity is not very low, we must take multiple reflection into consideration and treat the interference in thin films as a multibeam phenomenon rather than two-beam effect.

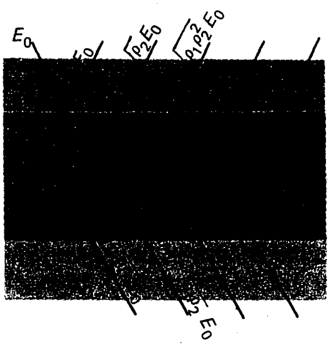


Fig. 137. To the calculation of reflectivity and transmission coefficient for a layer.

Figure 137 shows the multiple reflection in a plane-parallel plate of thickness d having a refractive index n_2, and abutting with media having refractive indices n_1 and n_3. The reflectivities of the upper and lower edges of the plate are denoted by ρ_1 and ρ_2 respectively. These quantities can be used to connect the amplitudes of the incident and reflected waves with the help of formula (18.5). The amplitudes of the incident and refracted waves are connected through the transmission coefficient τ in formula (18.9). These formulas can be written in a more convenient form as follows:

$$|E^0_{\mathrm{refr}}| = \sqrt{\tau v_2 \cos\theta_{\mathrm{in}}/(v_1\cos\theta_{\mathrm{refr}})}\,|E^0_{\mathrm{in}}| = \sqrt{\sigma}\,|E^0_{\mathrm{in}}|, \tag{29.10}$$

where

$$\sigma = \tau v_2 \cos\theta_{\mathrm{in}}/(v_1\cos\theta_{\mathrm{refr}}). \tag{29.11}$$

The reflectivity ρ depends on the angle of incidence. For the same angle of incidence, the reflectivity is generally different when waves from different media are incident at the interface. It can be seen from formulas (18.6) that for a certain angle of incidence θ_{in}, the reflectivity for a wave incidence from the first medium on the interface can be equal to the reflectivity for the wave incidence on the interface from the second medium only when the angle of incidence in the second case is equal to the angle of refraction in the first case. However, it can be seen from (16.30a) and (16.42a) that $\sqrt{\rho}$ has different signs when waves from different directions move to the interfaces. In other words, $\sqrt{\rho}$ must be taken equal to the right-hand sides of Eqs. (16.30a) and (16.42a) for calculating the refraction for the corresponding components of the electric field strength E. As regards σ [see (29.11)], it has different values for refraction at the same boundary under the above-mentioned conditions when waves propagating from different sides are considered. Hence we denote by σ_1 the coefficient σ for refraction at the upper surface of the plate when a ray of light is moving from the medium into the plate, and by σ_1' the corresponding coefficient for a ray moving from the plate into the medium. The coefficient for refraction at the lower surface of the plate when the ray of light moves from the plate into the medium is denoted by σ_2.

The amplitudes of rays transmitted and reflected at the plate surfaces are obtained from $\sqrt{\rho_1}$, $\sqrt{\rho_2}$, $\sqrt{\sigma_1}\sqrt{\sigma_1'}$ and $\sqrt{\sigma_2}$ by using the rules described above. On the basis of (29.4), we can write the phase difference between adjacent rays emerging from the plate as follows:

$$\delta = (4\pi/\lambda)\,n_2 d\cos\theta_{\mathrm{refr}}. \tag{29.12}$$

It is easier to calculate first the total complex amplitude of the intensity of the wave transmitted through the plate. This quantity is given by (see Fig. 137)

$$\begin{aligned} E &= \sqrt{\sigma_1\sigma_2}\,E_0 + \sqrt{\sigma_1\sigma_2}\,\sqrt{\rho_1\rho_2}\,E_0 e^{i\delta} \\ &+ \sqrt{\sigma_1\sigma_2}(\sqrt{\rho_1\rho_2})^2\,E_0 e^{2i\delta} + \dots \\ &= \sqrt{\sigma_1\sigma_2}\,E_0 \sum_{m=0}^{\infty} (\sqrt{\rho_1\rho_2})^m e^{mi\delta} \\ &= \sqrt{\sigma_1\sigma_2}\,E_0/(1 - \sqrt{\rho_1\rho_2}\,e^{i\delta}), \end{aligned} \tag{29.13}$$

where the sum of the geometric progression is extended to infinity since an infinite number of reflections contribute towards the interference in the case of a plane wave that is not bounded in space. For spatially bounded light beams, the number of reflections contributing to the interference observed through a finite-sized aperture is limited. However, even in this case the summation can be extended to an infinite number of reflections since the main contribution to the interference comes from the first few terms and the contribution from the remaining terms is negligible. Taking the squares of the moduli of both sides of (29.13), we obtain

$$|E|^2 = \frac{\sigma_1\sigma_2 |E_0|^2}{1 + \rho_1\rho_2 - 2\sqrt{\rho_1\rho_2}\cos\delta}. \tag{29.14}$$

In order to correctly apply (29.13) and the subsequent formulas obtained from it, it is necessary to emphasize the meaning of the quantities $\sqrt{\rho_1}$ and $\sqrt{\rho_2}$. From the derivation of formula (29.13), it is obvious that $\sqrt{\rho_1}$ characterizes the reflection of waves moving in the plate and reflected at the interface with the first medium back into the plate, while $\sqrt{\rho_2}$ corresponds to the reflection of a wave moving in the plate at the interface with the third medium. In other words, $\sqrt{\rho_1}$ and $\sqrt{\rho_2}$ describe the reflection of the waves in the plate at its surfaces.

The flux density of electromagnetic energy in the medium is

$$\boxed{S = EH = E^2/(\mu_0 v),} \tag{29.15}$$

where $\mu = \mu_0$ (the medium is nonmagnetic), and v is the velocity of propagation of the wave in the medium. The normal components of the flux of the wave energy from the first medium into the plate and from the plate into the third medium can be written in the form

$$S_{1n} \propto |E_0|^2 \cos\theta_{\text{in}\,1}/v_1, \tag{29.16}$$

$$S_{3n} \propto |E|^2 \cos\theta_{\text{refr}\,3}/v_3, \tag{29.17}$$

where $\theta_{\text{in}\,1}$ and $\theta_{\text{refr}\,3}$ represent the angle of incidence from the first medium to the plate surface and the angle of refraction from the plate to the third medium respectively.

Denoting by T the transmission coefficient of the plate as a whole, we can write

$$T = S_{3n}/S_{1n} = |E|^2 v_1 \cos\theta_{\text{refr}\,3}/(|E_0|^2 v_3 \cos\theta_{\text{in}\,1}). \tag{29.18}$$

Substituting $|E|^2/|E_0|^2$ from (29.14) into (29.18), we obtain

$$T = \frac{\sigma_1\sigma_2 v_1 \cos\theta_{\text{refr}\,3}/(v_3\cos\theta_{\text{in}\,1})}{1+\rho_1\rho_2 - 2\sqrt{\rho_1\rho_2}\cos\delta}. \tag{29.19}$$

Let us write σ_1 and σ_2 in accordance with their definition [see (29.11)]:

$$\sigma_1 = \tau_1 v_2 \cos\theta_{\text{in}\,1}/(v_1\cos\theta_{\text{refr}\,2});$$

$$\sigma_2 = \tau_2 v_3 \cos\theta_{\text{in}\,2}/(v_2\cos\theta_{\text{refr}\,3}). \tag{29.20}$$

Considering that $\cos\theta_{\text{refr}\,2} = \cos\theta_{\text{in}\,2}$, we get

$$\sigma_1\sigma_2 v_1 \cos\theta_{\text{refr}\,3}/(v_3\cos\theta_{\text{in}\,1}) = \tau_1\tau_2. \tag{29.21}$$

In this case, (29.19) assumes the form

$$T = \tau_1\tau_2/(1+\sqrt{\rho_1\rho_2} - 2\sqrt{\rho_1\rho_2}\cos\delta). \tag{29.22}$$

The law of energy conservation is satisfied for each surface:

$$\tau_1 = 1-\rho_1, \quad \tau_2 = 1-\rho_2. \tag{29.23}$$

If there is no absorption, the reflection coefficient for the entire plate can be written in the form

$$R = 1 - T = (\rho_1+\rho_2-2\sqrt{\rho_1\rho_2}\cos\delta)/(1+\rho_1\rho_2 - 2\sqrt{\rho_1\rho_2}\cos\delta), \tag{29.24}$$

where we have taken into consideration equalities (29.23).

!

Uniform inclination fringes are localized at infinity, while isopachic fringes are localized at the surface of the plate.
The finite size of the source does not blur the interference pattern of uniform inclination fringes and is not a factor limiting interference.
The admissible thickness of the films for observing interference in white light is limited by the capability of human eye to distinguish between close colours.

LAYER WITH ZERO REFLECTIVITY. If $R = 0$, there is no reflection from a layer. For this, the following condition must be satisfied

$$\rho_1+\rho_2-2\sqrt{\rho_1\rho_2}\cos\delta = 0. \tag{29.25}$$

We shall consider the normal incidence and choose a layer whose optical thickness is equal to one-fourth of a wavelength. In other words, we put $n_2 d = \lambda/4$, $\delta = \pi$ and $\cos\delta = -1$. Formula (29.24) then assumes the form

$$R = (\sqrt{\rho_1}+\sqrt{\rho_2})^2/(1+\sqrt{\rho_1}\sqrt{\rho_2})^2. \tag{29.26}$$

The coefficients $\sqrt{\rho_1}$ and $\sqrt{\rho_2}$ describe the reflection of a wave propagating in the plate (medium *2*) from the boundaries with media *1* and *3* respectively. Hence, on the basis of (16.33a), these coefficients are given by

$$\sqrt{\rho_1} = (n_2 - n_1)/(n_2 + n_1), \quad \sqrt{\rho_2} = (n_2 - n_3)/(n_2 + n_3). \qquad (29.27)$$

Substituting (29.27) into (29.26), we obtain

$$R = [(n_2^2 - n_1 n_3)/(n_2^2 + n_1 n_3)]^2. \qquad (29.28)$$

The reflection coefficient will be equal to zero if the following condition is satisfied:

$$\boxed{n_2 = \sqrt{n_1 n_3}\,.} \qquad (29.29)$$

For example, barium flint glass BF-1 has a refractive index $n_3 = 1.53$ for a wavelength 589.3 nm. In order to get rid of the reflection at the surface of this glass in air ($n_1 = 1$), we must coat the surface with a material whose refractive index is $n_2 = \sqrt{1.53} \simeq 1.23$ and whose optical thickness is equal to a quarter of the wavelength. The refractive index of most solids lies between 1.5 and 2.2, and the choice of a material with such a low refractive index is not a simple problem. This requirement is satisfied quite well by NaF whose refractive index is 1.33, and cryolite (Na_3AlF_6) which has nearly the same refractive index. The reflectivity in this case assumes the value 0.008 which is considerably lower than the value 0.04 of reflectivity at the flint glass surface. The reflectivity for other wavelengths is somewhat higher, but its value remains quite low over the entire optical range.

This method of decreasing the reflectivity of glass surfaces is called antireflection coating application.

LAYER WITH A HIGH REFLECTIVITY. The physical reason behind the low reflectivity of a layer whose optical thickness is a quarter of the wavelength and whose refractive index is lower than that of the material on which it is applied can be explained as follows. The phase of a wave reflected from both surfaces of the layer changes by π, and hence the phase difference between the reflected waves is due solely to two passages through a plate whose thickness is a quarter of the wavelength, and is thus equal to π. Hence the reflected waves suppress one another. In order to increase the reflectivity, we must ensure an augmentation of the waves reflected from different surfaces of the layer. This can be done with the help of a layer whose thickness is equal to a quarter of the wavelength if the phase of E changes at one of the surfaces by π, while the reflection at the other surface takes place without any phase change. For this purpose, the refractive index of the layer must

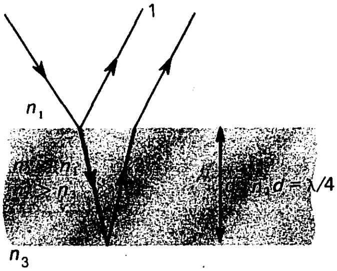


Fig. 138. Conditions for obtaining high reflectivity of a layer.

be higher or lower than the refractive indices of the adjoining media (Fig. 138). The phase of E changes by π as a result of reflection of the wave represented by the ray *1*. For the wave shown by the ray *2*, there is no phase change due to reflection. Hence the total phase difference between the reflected rays is equal to 2π and they augment each other. This leads to an increase in the reflectivity. For example, if the surface of glass ($n_3 = 1.5$) is coated with a layer of SiO_2 ($n_2 = 2$) having an optical thickness equal to a quarter of the wavelength, we obtain $R = 0.2$ from formula (29.28). This is much higher than the value $R = 0.04$ for reflection from the glass surface.

When both sides of a plate are abutted with the same medium, i.e. when $n_1 = n_3 = n$, formula (29.24) becomes simplified. In this case, obviously $\rho_1 = \rho_2 = \rho$ and the formula assumes the form

$$R = 2\rho(1 - \cos\delta)/(1 + \rho^2 - 2\rho\cos\delta). \tag{29.30}$$

If $\cos\delta = 1$, there is no reflection. This is possible when $\delta = 2\pi m$ ($m = 1, 2, \ldots$). For normal incidence, this condition assumes the form $n_2 d = m\lambda/2$. The highest reflection coefficient is obtained for $\cos\delta = -1$, which is the case when $\delta = (2m + 1)\pi$ ($m = 0, 1, 2, \ldots$). For normal incidence, this condition assumes the form $n_2 d = (2m + 1)\lambda/4$. The maximum reflection coefficient is

$$R_{\max} = 4\rho/(1 + \rho)^2. \tag{29.31}$$

For example, the reflectivity ρ of the glass surface ($n = 1.5$) in air ($n_1 = n_2 = 1$) is 0.04, while in accordance with formula (29.31), the maximum reflection coefficient $R_{\max} = 0.15$ for the case when reflection takes place at a glass plate in air. This is about four times higher than the value for reflection at one surface.

However, all that has been stated above is valid only for thin plates, when the optical path length in the plate is less than the coherence length. For example, in the case of isolated spectral lines from the ordinary radiation sources, where collision broadening as well as Doppler broadening take place, the plate thickness must be less than 1 mm. If laser radiation is used, the plate thickness may be several centimetres depending on the degree of coherence of laser radiation.

MATRIX METHOD OF COMPUTATIONS FOR MULTILAYER FILMS. Determination of the reflection coefficient for a film consisting of several layers is a very cumbersome problem. The easiest way to solve this problem is by using the matrix method which we describe here without derivation.

We assign the subscript 0 to quantities pertaining to the medium from which light falls on the film. Individual layers of

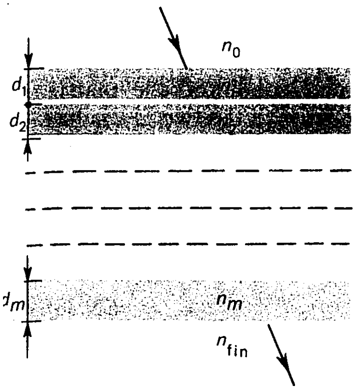


Fig. 139. Notation used in the matrix method of computations for multilayer films.

the film are assigned subscripts 1, 2, 3, ..., m. For the medium into which the light emerges from the film, we use the subscript "fin" (Fig. 139). Each row has a two-row square matrix corresponding to it:

$$M_j = \begin{pmatrix} \cos\delta_j & i\sin\delta_j/n_j \\ in_j\sin\delta_j & \cos\delta_j \end{pmatrix}, \tag{29.32}$$

where

$$\delta_j = (2\pi/\lambda)\, n_j d_j \cos\theta_{\text{refr}}, \tag{29.33}$$

n_j, d_j, $\cos\theta_{\text{refr}}$ are the refractive index, thickness and the angle of refraction for the j-th layer of the film. It should be noted that δ_j in (29.33) is just half the value of δ in (29.12).

The medium from which light is incident on the film is described by the two-row matrix

$$\begin{pmatrix} n_0 & -1 \\ n_0 & 1 \end{pmatrix}, \tag{29.34}$$

while the medium into which light emerges from the film is described by a column matrix

$$\begin{pmatrix} 1 \\ n_{\text{fin}} \end{pmatrix}. \tag{29.35}$$

The reflection coefficient is calculated from the formula

$$R = |a/b|^2, \tag{29.36}$$

where

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} n_0 & -1 \\ n_0 & 1 \end{pmatrix} \begin{pmatrix} \cos\delta_1 & \dfrac{i\sin\delta_1}{n_1} \\ in_1\sin\delta_1 & \cos\delta_1 \end{pmatrix}$$

$$\dots \begin{pmatrix} \cos\delta_m & \dfrac{i\sin\delta_m}{n_m} \\ in_m\sin\delta_m & \cos\delta_m \end{pmatrix} \begin{pmatrix} 1 \\ n_{\text{fin}} \end{pmatrix}. \tag{29.37}$$

MULTILAYER DIELECTRIC MIRRORS. Many optical instruments require reflecting surfaces with the highest possible reflection coefficient, say, 0.99 or even more. Moreover, it is often necessary that there should be the lowest possible absorption of light that is not reflected by this surface. Metallic mirrors do not satisfy the last condition since they strongly absorb the part of light that is not reflected by them. Multilayer dielectric mirrors are employed to satisfy this requirement.

A dielectric mirror element is formed by a couple of layers each having an optical thickness equal to a quarter of the wave-

?

How is a layer with zero reflectivity created?
Why does a layer whose optical thickness is equal to the quarter of a wavelength behave like a zero-reflectivity layer in some cases and has a very high reflectivity in some other cases?
How are multilayer dielectric mirrors with high reflectivity formed?

length. One of the layers is made of a material with a large refractive index n_l, while the other has a small refractive index n_s. Since $\delta_l = \delta_s = \pi/2$ by stipulated condition, the matrix describing the transmission of light through the two layers has the form

$$\begin{pmatrix} 0 & -i/n_l \\ -in_l & 0 \end{pmatrix}\begin{pmatrix} 0 & -i/n_s \\ -in_s & 0 \end{pmatrix} = \begin{pmatrix} -n_s/n_l & 0 \\ 0 & -n_l/n_s \end{pmatrix}. \tag{29.38}$$

The matrix describing the sequence of N such pairs is given by the expression

$$\begin{pmatrix} -n_s/n_l & 0 \\ 0 & -n_l/n_s \end{pmatrix}\cdots\begin{pmatrix} -n_s/n_l & 0 \\ 0 & -n_l/n_s \end{pmatrix} = \begin{pmatrix} (-n_s/n_l)^N & 0 \\ 0 & (-n_l/n_s)^N \end{pmatrix}. \tag{29.39}$$

Hence, in accordance with (29.37), we have

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} (-n_s/n_l)^N & 0 \\ 0 & (-n_l/n_s)^N \end{pmatrix}\begin{pmatrix} 1 \\ n_{\text{fin}} \end{pmatrix}. \tag{29.40}$$

Consequently, we obtain from formula (29.36)

$$R = \left[\frac{n_{\text{fin}} - (n_s/n_l)^{2N}}{n_{\text{fin}} + (n_s/n_l)^{2N}}\right]^2. \tag{29.41}$$

Since $n_s/n_l < 1$, the reflection coefficient tends to unity as N increases. In principle, this factor can be made indefinitely close to unity for a sufficiently large value of N. The number of pairs is limited by absorption and refraction of light as it passes through the layers. Values of R exceeding 0.995 have been attained at present.

TRANSLUCENT MATERIALS. By a suitable choice of the material of a plate and thickness of its layers, we can ensure that it has quite large values of the reflection and transmission coefficients (say, $R = 0.5$, $T = 0.5$). If such a material is used for manufacturing the window panes of passenger cars, it would be impossible to see in daylight what is happening inside the car since the light flux reflected from the glass panes will be much higher than the flux transmitted through the glass from the inside. On the other hand, all that is happening outside can be easily seen from inside.

SEC. 30. PARTIAL COHERENCE AND PARTIAL POLARIZATION

The general theory of partial coherence and partial polarization is described. The connection between this theory and the general theory of random processes is discussed.

PARTIAL COHERENCE. In the analysis of two-beam interference caused by amplitude division (see Sec. 26), it was found that the visibility of the interference pattern for strictly monochromatic waves is equal to unity. For quasi-monochromatic radiation, the visibility deteriorates with increasing path difference for the rays and ultimately becomes zero for a sufficiently large path difference exceeding the "coherence time length". When the visibility lies between 0 and 1, the waves are said to be partially coherent.

An identical situation holds for the case of two-beam interference caused by the division of the wave front (see Sec. 27). If the slits are illuminated by a monochromatic point source, the visibility of the interference pattern is equal to unity and the secondary waves emanating from the slits are said to be coherent. However, if the monochromatic source is not a point source, or if the point source is not monochromatic, the visibility of the interference pattern deteriorates. In both cases, the interfering waves are either partially coherent, or incoherent (if the interference pattern completely vanishes).

The two situations described above (see Secs. 26, 27) correspond to time coherence and space coherence respectively. The time coherence is characterized by the coherence length and time, while the space coherence is characterized by the coherence angle, width and radius. We can also introduce the concept of coherence volume. All these concepts express the idea of partial coherence which must be assigned a general quantitative expression.

MUTUAL COHERENCE FUNCTION. Figure 140*a* shows schematically the experimental set-up for obtaining an interference pattern. The rays emanating from points P_1 and P_2 interfere at point P_0. The paths of the rays from P_1 and P_2 to P_0 are shown by broken lines in order to emphasize the possibility of controlling their motion with the help of mirrors, lenses or other devices. To avoid complication of analysis for the sake of insignificant refinements, we shall assume that l_1 and l_2 are the path lengths and the velocity of light propagating along these paths is c. Hence the time consumed by the rays in passing along paths l_1 and l_2 is $t_1 = l_1/c$ and $t_2 = l_2/c$ respectively.

We denote the electric field strength at points P_1 and P_2 by $E_1(P_1, t')$ and $E_2(P_2, t')$ respectively.

The amplitudes of waves arriving at point P_0 from P_1 and P_2 vary. These variations depend on the path traversed by the waves and the size of the slits or other special devices

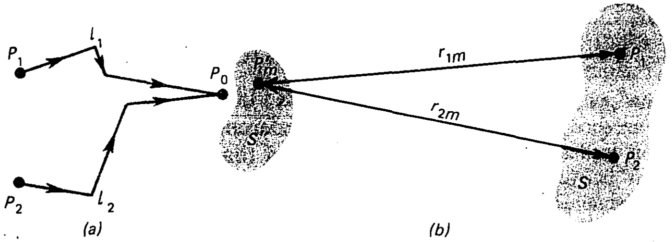


Fig. 140. To the determination of the coherence function (*a*); to the proof of the Van-Zittert-Zernike theorem (*b*).

employed for carrying out diffraction experiment. However, they do not affect the nature of variation of field strength with time. Hence the field strengths at point P_0 are obtained from the field strengths at points P_1 and P_2 by multiplying these values by certain constants taking into account the factors enumerated above. These constants include the factor $-i$ if $E_1(P_1, t')$ and $E_2(P_2, t')$ are the fields at points P_1 and P_2 and the waves propagating towards P_0 are considered as waves from secondary sources (see Sec. 32). In addition, we must take into account the time taken to traverse paths l_1 and l_2. Consequently, the total field strength at point P_0 is given by

$$E(P_0, t) = \alpha_1 E_1(P_1, t - l_1/c) + \alpha_2 E_2(P_2, t - l_2/c), \tag{30.1}$$

where α_1 and α_2 are constants defining the change in the wave characteristics upon passing from points P_1 and P_2 to point P_0, while the arguments $t - l_1/c$ and $t - l_2/c$ describe the time lags for the waves.

The intensity distribution in the interference pattern is described by the formula

$$I(P_0) = \langle \operatorname{Re} E \operatorname{Re} E \rangle = 1/2|\alpha_1|^2 \operatorname{Re}\langle |E_1|^2 \rangle \\ + 1/2|\alpha_2|^2 \operatorname{Re}\langle |E_2|^2 \rangle + |\alpha_1\alpha_2| \operatorname{Re}\langle E_1 E_2^* \rangle, \tag{30.2}$$

where for the sake of convenience of notation the arguments of the functions E_1 and E_2 have been omitted, it being clear that the arguments here are the same as in (30.1). It is assumed in (30.2) that α_1 and α_2 are either both real or both imaginary and hence $\alpha_1^*\alpha_2^* = \alpha_1\alpha_2$.

We shall consider the steady-state processes when the average quantities do not depend on the instant of time about which the averaging is carried out. Consequently,

$$1/2\langle E_1(P_1, t - l_1/c) E_1^*(P_1, t - l_1/c) \rangle \\ = 1/2\langle E_1(P_1, t) E_1^*(P_1, t) \rangle = I_1(P_1), \tag{30.3}$$

$$1/2\langle E_2(P_2, t - l_2/c) E_2^*(P_2, t - l_2/c) \rangle = I_2(P_2), \tag{30.4}$$

$$\langle E_1(P_1, t - l_1/c) E_2^*(P_2, t - l_2/c) \rangle \\ = \langle E_1(P_1, t) E_2^*(P_2, t + \tau) \rangle = \Gamma_{12}(\tau), \tag{30.5}$$

where $\tau = (l_1 - l_2)/v$ and $I_1(P_1)$, $I_2(P_2)$ are the wave intensities at points P_1 and P_2. **The function $\Gamma_{12}(\tau)$ is the basic quantity characterizing partial coherence. It describes the mutual coherence of optical vibrations at points P_1 and P_2 at instants of time separated by the interval τ, and is called the mutual coherence function of the wave field.** The points in the field to which this function corresponds are marked by letters with subscripts. If points P_1 and P_2 coincide ($P_1 = P_2$), the function

$$\Gamma_{11}(\tau) = \langle E_1(P_1, t) E_1^*(P_1, t + \tau)\rangle \tag{30.6}$$

is called the **self-coherence function of optical vibrations at point P_1**. In the general theory of random stationary processes, $\Gamma_{12}(\tau)$ is called the **mutual correlation function** for the quantities $E_1(t)$ and $E_2(t)$, while $\Gamma_{11}(\tau)$ is the **autocorrelation function** for the quantity $E_1(t)$.

Taking into account (30.3)-(30.5), we can present expression (30.2) in the form

$$I(P_0) = |\alpha_1|^2 I_1(P_1) + |\alpha_2|^2 I_2(P_2) + |\alpha_1\alpha_2| \operatorname{Re}\Gamma_{12}(\tau). \tag{30.7}$$

NORMALIZED COHERENCE FUNCTION. Considering that

$$|\alpha_1|^2 I_1(P_1) = I_1(P_0), \quad |\alpha_2|^2 I_2(P_2) = I_2(P_0),$$
$$|\alpha_1\alpha_2| = 2\sqrt{I_1(P_0)}\sqrt{I_2(P_0)}/\sqrt{\langle |E_1(P_1,t)|^2\rangle\langle |E_2(P_2,t)|^2\rangle}, \tag{30.8}$$

we can present (30.7) in the form

$$I(P_0) = I_1(P_0) + I_2(P_0) + 2\sqrt{I_1(P_0)}\sqrt{I_2(P_0)} \operatorname{Re}\gamma_{12}(\tau), \tag{30.9}$$

where

$$\gamma_{12}(\tau) = \Gamma_{12}(\tau)/\sqrt{\langle E_1(P_1,t)^2\rangle\langle |E_2(P_2,t)|^2\rangle} \tag{30.10}$$

is the **normalized coherence function.** Formula (30.9) expresses the general law of interference for stationary wave fields. This formula was derived earlier for particular cases [see, for example, (26.7)]. If the interference is caused by the amplitude division, say, in a Michelson interferometer, $\gamma_{12}(\tau)$ in formula (30.9) is replaced by $\gamma_{11}(\tau)$ corresponding to point P_1 at which the amplitude division takes place.

COHERENCE FUNCTION. The concept of coherence is related to the visibility of the interference pattern. Let us use formula (30.9) to determine the visibility of the interference pattern formed by the radiation from a quasi-monochromatic source with an average frequency ω. The wave fields at points

P_1 and P_2 have the following expressions:

$$E_1(P_1, t') = E_{01}(t')\exp\{-i[\omega t' - \varphi_1(t')]\}, \tag{30.11}$$

$$E_2(P_2, t') = E_{02}(t')\exp\{-i[\omega t' - \varphi_2(t')]\}, \tag{30.12}$$

where amplitudes E_1, E_2 and phases φ_1 and φ_2 are random functions, and E_{01}, E_{02} are assumed to be real. The normalized coherence function is given by the formula

$$\gamma_{12}(\tau) = \langle E_{01}(t)E_{02}(t+\tau)\exp \times \{i[\omega\tau - \psi(\tau)]\}\rangle / \sqrt{\langle |E_{01}|^2\rangle\langle E_{02}|^2\rangle}, \tag{30.13}$$

where

$$\psi = \varphi_2(t+\tau) - \varphi_1(t). \tag{30.14}$$

The factor $\exp(i\omega\tau)$ can be taken out of the averaging sign, and (30.13) can be transformed as follows:

$$\gamma_{12}(\tau) = (\langle E_{01}E_{02}\cos\psi\rangle - i\langle E_{01}E_{02}\sin\psi\rangle)\,e^{i\omega\tau} / \sqrt{\langle |E_{01}|^2\rangle\langle |E_{02}|^2\rangle} = |\gamma_{12}(\tau)|\,e^{-i\varphi}e^{i\omega\tau} = |\gamma_{12}(\tau)|\,e^{i(\omega\tau-\varphi)}, \tag{30.15}$$

where

$$\tan\varphi = \langle E_{01}E_{02}\sin\psi\rangle / \langle E_{01}E_{02}\cos\psi\rangle, \tag{30.16}$$

$$|\gamma_{12}(\tau)| = \sqrt{(\langle E_{01}E_{02}\cos\psi\rangle)^2 + (\langle E_{01}E_{02}\sin\psi\rangle)^2} / \sqrt{\langle |E_{01}|^2\rangle\langle |E_{02}|^2\rangle}. \tag{30.17}$$

Using Schwarz's inequality, we can easily show that $|\gamma_{12}(\tau)| \leqslant 1$. Formula (30.9) then assumes the form

$$\underline{\overline{I = I_1 + I_2 + 2\sqrt{I_1}, \ \sqrt{I_2}|\gamma_{12}(\tau)|\cos(\omega\tau - \varphi).}} \tag{30.18}$$

For quasi-monochromatic light, the quantities $|\gamma_{12}(\tau)|$ and φ are slowly varying functions in comparison with $\cos\omega\tau$, as can be seen from an analysis of interference patterns (see Secs 26, 27). Hence the intensity maxima and minima in (30.18) are attained for values of $\cos(\omega\tau - \varphi)$ equal to ± 1, and the visibility of the interference pattern is

$$\underline{\overline{V = 2\sqrt{I_1}\sqrt{I_2}|\gamma_{12}(\tau)|/(I_1 + I_2).}} \tag{30.19}$$

This formula directly connects the visibility of an interference pattern with the degree of coherence. For interfering beams with the same intensity ($I_1 = I_2$), the visibility is equal to the coherence function:

$$\underline{\overline{V = |\gamma_{12}(\tau)|.}} \tag{30.20}$$

!

The mutual coherence function of the wave field and the self-coherence function of optical vibrations in the general theory of random stationary processes are respectively called the mutual correlation function and autocorrelation function. The normalized coherence function contains information about the amplitude and phase fluctuations of the wave.
If the interfering beams of light have the same intensity, the visibility is equal to the coherence function.
The theory of partial polarization is based on an analysis of the mutual coherence function of the mutually perpendicular components of the electric field strength of the wave.

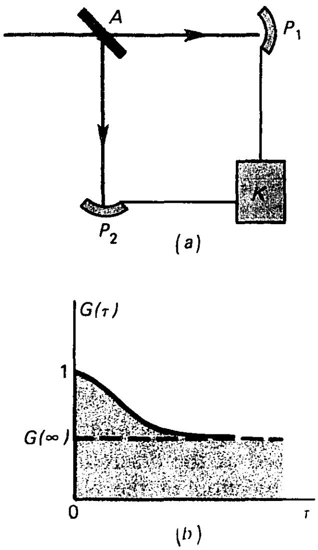


Fig. 141. Schematic diagram of the Brown-Twiss experiment (*a*); correlation of intensity vs τ (*b*).

BROWN-TWISS EXPERIMENT. This experiment was performed to study the correlation of intensity in a light beam. The light flux from source S (Fig. 141*a*) is divided into two parts by a translucent plate A. These split beams are directed towards photodetectors P_1 and P_2 and traverse different paths. The current from the detector, which is proportional to the light flux, is directed into correlator C where a current equal to the product of the currents is generated in appropriate electric circuits. The measurable quantity in this case is

$$G(\tau) = \frac{1}{\langle I^2 \rangle} \frac{1}{T} \int_0^T I(t) I(t+\tau)\,\mathrm{d}t. \tag{30.21}$$

Since $I \propto E^2$, we are dealing here with a fourth-order correlation function in electric field strength. The dependence of G on τ, determined in the experiments of Brown and Twiss, is shown in Fig. 141*b*. For very low τ, the value of $G(\tau)$ is close to unity. With increasing τ, this quantity decreases and attains a constant value for large τ.

To explain such a behaviour of $G(\tau)$, we must take into account the fluctuations in the intensity of the light beam. In the absence of fluctuations, we would have $G(\tau) = 1$ for all values of τ. However, the situation changes in the presence of fluctuations. A characteristic time scale can be defined for the fluctuations. If τ is less than the characteristic time, identical currents are registered in the correlator all the time and $G(\tau)$ is close to unity. As τ increases, the correlation between the currents in the correlator is violated, the current maxima corresponding to one circuit fall on the minima in the other circuit, and so on. Consequently, $G(\tau)$ decreases. If τ increases beyond the characteristic fluctuation time, its further increase does not cause any variation in the ratio of currents in the circuits and the value of $G(\tau)$ remains unchanged. The function $G(\tau)$ gives information about the statistical properties of radiation.

PARTIAL POLARIZATION. In a quasi-monochromatic wave, the tip of the electric field vector describes a highly irregular curve in the plane perpendicular to the direction of wave propagation (see Sec. 13). In a wave propagating along the positive direction of the Z-axis, the projections of the electric field strength at a fixed point can be written in a form analogous to (13.5):

$$E_x(t) = E_{0x}(t) \exp\{-i[\omega t - \varphi_x(t)]\}, \tag{30.22}$$

$$E_y(t) = E_{0y}(t) \exp\{-i[\omega t - \varphi_y(t)]\}, \tag{30.23}$$

where E_{0x}, E_{0y}, φ_x, φ_y are random amplitudes and phases. For a strictly monochromatic wave, these quantities are constant

?

What information is lost as we go over from normalized coherence function to the coherence function?

In what way is the visibility of the interference pattern related to the coherence function in the case of an arbitrary ratio of the intensities of the interfering beams?

What is the coherence matrix of a quasi-monochromatic wave?

Define the degree of polarization of a light wave and write down its expression in terms of the extremal values of intensity.

Write down the expression for natural light as a superposition of linearly and circularly polarized waves.

and the wave has quite a definite polarization. A quasi-monochromatic wave does not have a definite polarization. However, if a certain regularity is observed in the XY-plane in the vibrations of the electric field vector, we can speak of a partial polarization. The quantitative theory of partial polarization is based on an analysis of the correlation between the two mutually perpendicular projections of the electric field vector and is characterized by the coherence matrix.

COHERENCE MATRIX FOR A QUASI-MONOCHROMATIC PLANE WAVE. Let us calculate the intensity of optical vibrations in a direction forming an angle θ with the positive direction of the X-axis. The projection of the electric field vector onto this direction is given by the formula

$$E(t, \theta, \varepsilon) = E_x \cos\theta + E_y \sin\theta \, e^{i\varepsilon}, \tag{30.24}$$

where ε is the phase shift between the projections and is an independent parameter. Its value is fixed at the discretion of the experimenter. For $\varepsilon > 0$, the y-projection of the electric field strength of the wave lags behind the x-projection, while for $\varepsilon < 0$ the y-projection leads the x-projection. For the wave intensity, we obtain the following formula:

$$\begin{aligned} I(\theta, \varepsilon) &= \langle \operatorname{Re} E \operatorname{Re} E \rangle = (1/2) \operatorname{Re} \langle E E^* \rangle = (1/2) \cos^2\theta \langle E_x E_x^* \rangle \\ &+ (1/2) \sin^2\theta \langle E_y E_y^* \rangle + \sin\theta \cos\theta \operatorname{Re} \langle E_x E_y^* e^{-i\varepsilon} \rangle \\ &= I_{xx} \cos^2\theta + I_{yy} \sin^2\theta + 2 \sin\theta \cos\theta \operatorname{Re}(I_{xy} e^{-i\varepsilon}), \end{aligned} \tag{30.25}$$

where

$$\begin{aligned} I_{xx} &= (1/2)\langle E_x E_x^* \rangle, & I_{xy} &= (1/2)\langle E_x E_y^* \rangle, \\ I_{yx} &= (1/2)\langle E_y E_x^* \rangle, & I_{yy} &= (1/2)\langle E_y E_y^* \rangle. \end{aligned} \tag{30.26}$$

The matrix

$$J = \begin{Vmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{Vmatrix} \tag{30.27}$$

is called the **coherence matrix of a quasi-monochromatic plane wave.** It is a Hermitian matrix since $I_{xy} = I_{yx}^*$.

NORMALIZED COHERENCE FUNCTION FOR MUTUALLY PERPENDICULAR PROJECTIONS OF THE ELECTRIC FIELD STRENGTH OF A WAVE. In complete analogy with (30.13), the normalized coherence function μ_{xy} of two mutually perpendicular projections of the electric field strength of a wave is defined by the formula

$$\mu_{xy} = \langle E_x E_y^* \rangle / \sqrt{\langle E_x E_x^* \rangle \langle E_y E_y^* \rangle} = I_{xy} / \sqrt{I_{xx} I_{yy}}. \tag{30.28}$$

Taking this relation into account, we can write (30.25) in the form

$$I(\theta, \varepsilon) = I_{xx}\cos^2\theta + I_{yy}\sin^2\theta + 2\sin\theta\cos\theta \times \sqrt{I_{xx}}\sqrt{I_{yy}}\,\mathrm{Re}(\mu_{xy}e^{-i\varepsilon}), \tag{30.29}$$

which is analogous to (30.9).

For a quasi-monochromatic wave described by Eqs. (30.22) and (30.23), we have

$$I_{xy} = \langle E_{0x}E_{0y}e^{i\psi}\rangle = \langle E_{0x}E_{0y}\cos\psi\rangle + i\langle E_{0x}E_{0y}\sin\psi\rangle,$$

where $\psi = \varphi_x - \varphi_y$.

Consequently,

$$I_{xy} = |I_{xy}|e^{i\varphi},$$

$$|I_{xy}| = \sqrt{(\langle E_{0x}E_{0y}\cos\psi\rangle)^2 + (\langle E_{0x}E_{0y}\sin\psi\rangle)^2}, \tag{30.30}$$

where

$$\tan\varphi = \langle E_{0x}E_{0y}\sin\psi\rangle / \langle E_{0x}E_{0y}\cos\psi\rangle.$$

Hence the normalized coherence function can be written in the form

$$\mu_{xy} = |\mu_{xy}|e^{i\varphi}, \tag{30.31a}$$

where

$$|\mu_{xy}| = |I_{xy}|/\sqrt{I_{xx}I_{yy}} \tag{30.31b}$$

is the **coherence function of mutually perpendicular projections of the electric field strength of a wave.** Formula (30.29) then assumes the form

$$\begin{aligned} I(\theta, \varepsilon) &= I_{xx}\cos^2\theta + I_{yy}\sin^2\theta + 2\sin\theta\cos\theta \\ &\times \sqrt{I_{xx}}\sqrt{I_{yy}}\,|\mu_{xy}|\cos(\varepsilon - \varphi) \\ &= I_{xx}\cos^2\theta + I_{yy}\sin^2\theta + 2\sin\theta\cos\theta\sqrt{I_{xy}I_{yx}}\cos(\varepsilon - \varphi). \end{aligned} \tag{30.32}$$

The normalized coherence function μ_{xy} contains complete information about the mutual coherence of the x- and y-projections of the electric field strength of a wave. It must be noted that the coherence function $|\mu_{xy}|$ generally depends on the direction of the coordinate axes. The coherence function is equal to zero only for a completely nonpolarized light, irrespective of the direction of coordinate axes, in a plane perpendicular to the direction of wave propagation.

NATURAL (NONPOLARIZED) LIGHT. The intensity of natural light is the same in all directions and does not depend on the phase difference that may appear between two mutually perpendicular projections of the electric field strength of the wave. In other words, the intensity is independent of θ and ε, or $I(\theta, \varepsilon) = \text{const}$. It can be seen from (30.32) that this is possible only for

$$\mu_{xy} = 0, \quad I_{xy} = I_{yx} = 0, \tag{30.33}$$

$$I_{xx} = I_{yy} = (1/2)I, \tag{30.34}$$

where $I = I_{xx} + I_{yy}$ is the total intensity of light.

Thus, the coherence function for two mutually perpendicular projections of the electric field strength is equal to zero for nonpolarized light, while the coherence matrix has the form

$$J = \frac{1}{2}I \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}. \tag{30.35}$$

COMPLETELY POLARIZED LIGHT. To begin with, we shall consider monochromatic waves. These are described by formulas (30.22) and (30.23), where E_{0x}, E_{0y}, φ_x and φ_y are independent of time. In this case, the average values are equal to instantaneous values and the coherence matrix has the form

$$J = \frac{1}{2} \left\| \begin{matrix} E_{0x}^2 & E_{0x}E_{0y}e^{i\psi} \\ E_{0x}E_{0y}e^{-i\psi} & E_{0y}^2 \end{matrix} \right\|. \tag{30.36}$$

The quantity $\psi = \varphi_x - \varphi_y$ is the constant phase difference between the mutually perpendicular projections of the electric field strength of a wave. In accordance with (30.28), the normalized coherence function is given by the formula

$$\mu_{xy} = e^{i\psi}. \tag{30.37}$$

Consequently,

$$|\mu_{xy}| = 1. \tag{30.38}$$

This means that the coherence function varies from 0 to 1 as we go from natural light to completely polarized light. Values between 0 and 1 correspond to partial polarization. However, polarization is quantitatively characterized by the degree of polarization P [see (30.63)] which is not equal to $|\mu_{xy}|$.

The determinant of the coherence matrix (30.36) for completely polarized light is given by

$$\det J = I_{xx}I_{yy} - I_{xy}I_{yx} = 0. \tag{30.39}$$

This equality is the most important condition for complete polarization.

Complete polarization may be observed not only in monochromatic waves, but also in quasi-monochromatic waves for which condition (30.39) for complete polarization is satisfied if

$$E_{0y}(t)/E_{0x}(t) = \text{const} = \alpha, \quad \varphi_x - \varphi_y = \psi = \text{const}. \tag{30.40}$$

We have

$$I_{xx} = \langle E_{0x}^2 \rangle, \; I_{xy} = \alpha \langle E_{0x}^2 \rangle e^{i\psi},$$
$$I_{yx} = \alpha \langle E_{0x}^2 \rangle, \; I_{yy} = \alpha^2 \langle E_{0x}^2 \rangle.$$

Hence Eq. (30.39) is satisfied.

This conclusion can be also derived from physical considerations. Equation (30.39) shows that I_{xx}, I_{xy}, I_{yx} and I_{yy} are not independent for completely polarized light. Condition (30.40) expresses the dependence of projections E_{0x} on E_{0y} and their phases, which ensures that relation (30.39) for completely polarized light is satisfied.

Linearly polarized light is defined by the condition $\psi = m\pi$ ($m = 0, \pm 1, \pm 2, \ldots$). The coherence matrix (30.36) assumes the form

$$J = \frac{1}{2} \begin{Vmatrix} E_{0x}^2 & (-1)^m E_{0x} E_{0y} \\ (-1)^m E_{0x} E_{0y} & E_{0y}^2 \end{Vmatrix}. \tag{30.41}$$

The direction of oscillations of vector E is defined by the ratio $E_x/E_y = (-1)^m E_{0x}/E_{0y}$. If the electric vector oscillates in the direction of the X-axis ($E_{0y} = 0$), we obtain

$$J = I \begin{Vmatrix} 1 & 0 \\ 0 & 0 \end{Vmatrix}, \tag{30.42}$$

where $I = E_{0x}^2/2$ is the wave intensity. For oscillations along the Y-axis, we have

$$J = I \begin{Vmatrix} 0 & 0 \\ 0 & 1 \end{Vmatrix}. \tag{30.43}$$

If, however, the electric field vector of a wave oscillates along the bisectors of the angles between the coordinate axes, i.e. at angles $\pi/4$ and $3\pi/4$ to the X-axis, then $E_{0x} = E_{0y}$, $m = 0$ and $E_{0x} = E_{0y}$, $m = 1$. The coherence matrices then assume the form

$$J = \frac{1}{2} I \begin{Vmatrix} 1 & 1 \\ 1 & 1 \end{Vmatrix}, \qquad J = \frac{1}{2} I \begin{Vmatrix} 1 & -1 \\ -1 & 1 \end{Vmatrix}. \tag{30.44}$$

For circularly polarized light, $E_{0x} = E_{0y}$, $\psi = m\pi/2$ ($m = \pm 1, \pm 3, \ldots$) and hence the coherence matrix is written as follows:

$$J = \frac{1}{2} I \begin{Vmatrix} 1 & \pm i \\ \mp i & 1 \end{Vmatrix}, \tag{30.45}$$

where the upper and the lower signs of i refer to the opposite directions of circular polarization.

DEGREE OF POLARIZATION OF A LIGHT WAVE. There is no interference between independent light waves. Hence the mean values of the products of the projections of different waves are equal to zero. This leads to the conclusion that the coherence matrix of the resulting wave is equal to the sum of coherence matrices of the component waves. In order to prove this, let us consider the superposition of independent waves. The projections of the electric field strength of the resultant wave are given by the equalities.

$$E_x = \sum_{n=1}^{N} E_x^{(n)}, \quad E_y = \sum_{m=1}^{N} E_y^{(m)}. \tag{30.46}$$

Hence we obtain the following expression for the average values of the products of the projections:

$$\begin{aligned} \langle E_i E_j^* \rangle &= \sum_{n=1}^{N} \sum_{m=1}^{N} \langle E_i^{(n)} E_j^{(m)*} \rangle \\ &= \sum_{n=1}^{N} \langle E_i^{(n)} E_j^{(n)*} \rangle \sum_{n=m}^{N} \langle E_i^{(n)} E_j^{(m)*} \rangle, \end{aligned} \tag{30.47}$$

where $i = x, y$ and $j = x, y$. The terms in the last sum in this equation are equal to zero since they are the averages of the products of projections of electric field strengths of different independent waves. Hence we arrive at the equality

$$\langle E_i E_j^* \rangle = \sum_{n=1}^{N} \langle E_i^{(n)} E_j^{(n)*} \rangle, \tag{30.48}$$

which shows that the coherence matrix of the resultant wave is equal to the sum of the coherence matrices of the component waves. This suggests the possibility that the coherence matrix of any wave can be presented as the sum of coherence matrices of a completely polarized wave and a completely nonpolarized wave. To realize this possibility, we must prove the uniqueness of the matrix representation

$$J = \begin{Vmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{Vmatrix}. \tag{30.49}$$

in the form

$$J = J^{(1)} + J^{(2)}, \tag{30.50}$$

where

$$J^{(1)} = \begin{Vmatrix} A & 0 \\ 0 & A \end{Vmatrix} \tag{30.51}$$

is the coherence matrix of a completely nonpolarized light in accordance with (30.35), and

$$J^{(2)} = \begin{Vmatrix} B & D \\ D^* & C \end{Vmatrix}, \tag{30.52}$$

which satisfies condition (30.39) in the form

$$BC - DD^* = 0, \tag{30.53}$$

is the coherence matrix of a completely polarized light. From (30.50) we obtain the equalities

$$A + B = I_{xx}, \quad D = I_{xy}, \quad D^* = I_{yx}, \quad A + C = I_{yy} \tag{30.54}$$

which can be used to reduce Eq. (30.53) to the form

$$A^2 - (I_{xx} + I_{yy})A + I_{xx}I_{yy} - I_{xy}I_{yx} = 0. \tag{30.55}$$

The roots of this equation are

$$A = \frac{1}{2}(I_{xx} + I_{yy}) \pm \frac{1}{2}\sqrt{(I_{xx} + I_{yy})^2 - 4\det J}, \tag{30.56}$$

where

$$\det J = I_{xx}I_{yy} - I_{xy}I_{yx} \tag{30.57}$$

is the determinant of the matrix (30.49).

In order to analyze the solution of this equation, we must take into acount certain inequalities. From (30.28) we obtain

$$1 - \mu_{xy}\mu_{xy}^* = (I_{xx}I_{yy} - I_{xy}I_{yx})/(I_{xx}I_{yy}) = \det J/(I_{xx}I_{yy}), \tag{30.58}$$

$$\mu_{xy} = I_{xy}/\sqrt{I_{xx}I_{yy}}. \tag{30.59}$$

In accordance with Schwarz's inequality, it follows from (30.59) that $|\mu_{xy}| \leqslant 1$. Considering that I_{xx} and I_{yy} are nonnegative, we obtain from (30.58) $0 \leqslant \det J \leqslant I_{xx}I_{yy}$. Furthermore, from the inequality $(I_{xx} - I_{yy})^2 \geqslant 0$, we obtain $(I_{xx} + I_{yy})^2 \geqslant 4I_{xx}I_{yy}$, and hence $0 \leqslant \det J \leqslant I_{xx}I_{yy} \leqslant (1/4)(I_{xx} + I_{yy})^2$. This means that both roots in (30.56) are real and positive. A direct inspection shows that the positive sign of the root in (30.56) means that B and C are negative, which is unacceptable since B and C are positive by definition. Hence we must take the negative sign for the root in (30.56). The solution is unique and has the form

$$A = \frac{1}{2}(I_{xx} + I_{yy}) - \frac{1}{2}\sqrt{(I_{xx} + I_{yy})^2 - 4\det J}, \tag{30.60a}$$

$$B = \frac{1}{2}(I_{xx} - I_{yy}) + \frac{1}{2}\sqrt{(I_{xx} + I_{yy})^2 - 4\det J}, \tag{30.60b}$$

$$C = \frac{1}{2}(I_{yy} - I_{xx}) + \frac{1}{2}\sqrt{(I_{xx} + I_{yy})^2 - 4\det J}, \tag{30.60c}$$

$$D = I_{xy}, \quad D^* = I_{yx}. \tag{30.60d}$$

The total intensity of the wave and the intensity of the polarized part are given by

$$I = I_{xx} + I_{yy}, \tag{30.61}$$

$$I_{\text{tot}} = B + C = \sqrt{(I_{xx} + I_{yy})^2 - 4\det J}. \tag{30.62}$$

The degree of polarization P of a wave is the ratio of the intensity of the polarized part to the total intensity:

$$\boxed{P = \frac{I_{\text{tot}}}{I} = \sqrt{1 - 4\det J/(I_{xx} + I_{yy})^2}.} \tag{30.63}$$

From the inequalities considered for the solution of (30.60) it follows that $0 \leqslant P \leqslant 1$. It is well known from the matrix theory that $\det J$ and $\operatorname{Tr} J = I_{xx} + I_{yy}$ are invariant to the rotation of the coordinate system. Hence, taking into account (30.63), we conclude that P is independent of the direction of X- and Y-axes.

EXPRESSION OF DEGREE OF POLARIZATION IN TERMS OF EXTREMAL INTENSITIES. Let us study the intensity as a function of θ and ε. Using the trigonometric formulas for the sines and cosines of double angles, we can write (30.32) in the form

$$I(\theta, \varepsilon) = (1/2)(I_{xx} + I_{yy}) + (1/2)(I_{yy} - I_{xx})\cos 2\theta$$
$$+ \sqrt{I_{xy}I_{yx}}\cos(\varepsilon - \varphi)\sin 2\theta. \tag{30.64}$$

This formula can be transformed in the same way as used for transforming (26.23) into (26.25a):

$$I(\theta, \varepsilon) = Q + \sqrt{C^2 + S^2}\cos(2\theta - \alpha), \tag{30.65}$$

where

$$Q = (I_{xx} + I_{yy})/2, \quad C = (I_{yy} - I_{xx})/2,$$
$$S = \sqrt{I_{xy}I_{yx}}\cos(\varepsilon - \varphi), \quad \tan\alpha = S/C.$$

The amplitude of intensity oscillations depends on $\cos(\varepsilon - \varphi)$ and attains its maximum value for $|\cos(\varepsilon - \varphi)| = 1$. The maxima and minima of intensity are attained at angles θ for which $\cos(2\theta - \alpha) = 1$ and $\cos(2\theta - \alpha) = -1$ respectively. Consequently, we get

$$I(\theta, \varepsilon)|_{\max} = Q + \sqrt{C^2 + S^2},$$
$$I(\theta, \varepsilon)|_{\min} = Q - \sqrt{C^2 + S^2}.$$

The methods for determining the extrema described above must be used with care. As a matter of fact, the dependence of I in (30.65) on θ and ε appears not only in the arguments $\cos(\varepsilon - \varphi)$ and $\cos(2\theta - \alpha)$, but also in the quantities I_{xx}, I_{xy}, I_{yx} and I_{yy}. The use of the method described above is justified by the fact that the dependence of I on ε and θ through $\cos(\varepsilon - \varphi)$ and $\cos(2\theta - \alpha)$ is much stronger than the dependence I_{xx}, I_{xy}, I_{yx} and I_{yy}. The latter quantities can, therefore, be neglected while determining the intensity extrema.

Considering that

$$\sqrt{C^2 + S^2} = \frac{1}{2}\sqrt{(I_{yy} - I_{xx})^2 + 4I_{xy}I_{yx}}$$
$$= \frac{1}{2}\sqrt{(I_{yy} + I_{xx})^2 - 4(I_{xx}I_{yy} - I_{xy}I_{yx})}$$
$$= \frac{1}{2}\sqrt{(I_{yy} + I_{xx})^2 - 4\det J},$$

where $\det J = I_{xx}I_{yy} - I_{xy}I_{yx}$ is the determinant of the coherence matrix; we obtain

$$\frac{I(\theta, \varepsilon)_{\max} - I(\theta, \varepsilon)_{\min}}{I(\theta, \varepsilon)_{\max} + I(\theta, \varepsilon)_{\min}} = \frac{\sqrt{C^2 + S^2}}{Q} = \sqrt{1 - 4\det J/(I_{xx} + I_{yy})^2}. \tag{30.66}$$

A comparison of the right-hand sides of (30.66) and (30.63) gives

$$P = (I(\theta, \varepsilon)|_{\max} - I(\theta, \varepsilon)|_{\min})/(I(\theta, \varepsilon)|_{\max} + I(\theta, \varepsilon)|_{\min}\,. \tag{30.67}$$

By the extremal values in this expression we mean the absolute maxima and minima with respect to both independent parameters θ and ε. Completely polarized light is characterized by the value $P = 1$. If follows from (30.63) that in this case $\det J = I_{xx}I_{yy} - I_{xy}I_{yx} = 0$. This means that $|\mu_{xy}| = 1$, i.e. E_x and E_y are completely coherent with respect to each other. Completely nonpolarized light is characterized by the value

$P = 0$. It follows from (30.63) that

$$(I_{xx} + I_{yy})^2 - 4(I_{xx}I_{yy} - I_{xy}I_{yx}) = (I_{xx} - I_{yy})^2 + 4I_{xy}I_{yx} = 0. \tag{30.68}$$

Since $I_{xy}I_{yx} = I_{xy}I^*_{xy}$ is a positive quantity, each term on the right-hand side of (30.68) must be separately equal to zero:

$$I_{xx} = I_{yy}, \quad I_{xy} = I_{yx} = 0. \tag{30.69}$$

Consequently, $|\mu_{xy}| = 0$, and E_x and E_y are completely incoherent with respect to each other. For $0 < P < 1$, the light is partially polarized, $0 < |\mu_{xy}| < 1$, and E_x, E_y are partially coherent with respect to each other.

REPRESENTATION OF NATURAL LIGHT. Using the representation of the coherence matrix for a resultant wave as the sum of coherence matrices of the component waves, coherence matrix (30.35) for the natural light can be presented in the form

$$\frac{1}{2}I\left\|\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}\right\| = \frac{1}{2}I\left\|\begin{matrix} 1 & 0 \\ 0 & 0 \end{matrix}\right\| + \frac{1}{2}I\left\|\begin{matrix} 0 & 0 \\ 0 & 1 \end{matrix}\right\|, \tag{30.70}$$

$$\frac{1}{2}I\left\|\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}\right\| = \frac{1}{4}I\left\|\begin{matrix} 1 & +i \\ -i & 1 \end{matrix}\right\| + \frac{1}{4}I\left\|\begin{matrix} 1 & -i \\ +i & 1 \end{matrix}\right\|. \tag{30.71}$$

In the first case, a natural light wave of intensity I is presented as the sum of two linearly polarized waves with intensities $I/2$. The electric vectors oscillate in mutually perpendicular directions in a plane perpendicular to the direction of wave propagation. In the second case, a natural light wave is equivalent to two independent waves with left and right circular polarizations. The intensity of each component wave is equal to half the intensity I of natural light.

RELATION BETWEEN THE DEGREES OF POLARIZATION AND COHERENCE FUNCTION. From (30.59), we obtain

$$|\mu_x|^2 = 1 - \det J/(I_{xx}I_{yy}). \tag{30.72}$$

The square of the degree of polarization (30.63) is expressed through the formula

$$P^2 = 1 - \det J/[(I_{xx} + I_{yy})^2/4]. \tag{30.73}$$

It follows from the inequality $(I_{xx} - I_{yy})^2 \geqslant 0$ that $(I_{xx} + I_{yy})^2 \geqslant 4I_{xx}I_{yy}$. Hence a comparison of (30.72) and (30.73) shows that

$$P \geqslant |\mu_{xy}|. \tag{30.74}$$

The equality is attained only for $I_{xx} = I_{yy}$. To find out when this situation is realized, we consider the transformation of the projections of electric field strengths as a result of a rotation of coordinate axes. If the X'-axis of the new system of coordinates forms an angle α with the X-axis (the two systems have the same origin and the rotation is carried out in the XY-plane), we obtain

$$E_{x'} = E_x \cos\alpha + E_y \sin\alpha, \quad E_{y'} = -E_x \sin\alpha + E_y \cos\alpha. \tag{30.75}$$

In this case,

$$\begin{aligned} I_{x'x'} &= I_{xx}\cos^2\alpha + I_{yy}\sin^2\alpha + (I_{xy} + I_{yx})\cos\alpha\sin\alpha, \\ I_{y'y'} &= I_{xx}\sin^2\alpha + I_{yy}\cos^2\alpha - (I_{xy} + I_{yx})\sin\alpha\cos\alpha. \end{aligned} \tag{30.76}$$

From the condition $I_{x'x'} = I_{y'y'}$, we get

$$\tan 2\alpha = (I_{yy} - I_{xx})/(I_{xy} + I_{yx}). \tag{30.77}$$

Since $I_{xy} = I_{yx}$, the solution of Eq. (30.77) gives two mutually orthogonal directions for which the intensities $I_{x'x'}$ and $I_{y'y'}$, are identical and the coherence function $\mu_{x'y'}$ attains its maximum value equal to the degree of polarization P of the wave.

VAN-ZITTERT-ZERNIKE THEOREM. In Secs. 26 and 27, we considered some specific cases in which time and space coherence are manifested. Since coherence determines the visibility of the interference pattern, we must know how to determine the coherence function for radiation without knowing the visibility of the interference pattern. For a quasi-monochromatic radiation from a small-size source located at a large distance, such a possibility is presented by the **Van-Zittert-Zernike theorem. This theorem is valid** not for any small source at a sufficiently large distance but only **under the condition that the difference in the propagation times for light from any point of the source up to any two points under observation, for which the mutual coherence of radiation is being investigated, is smaller than the coherence time for the radiation from the corresponding points on the source.**

To simplify the calculations, we assume that the **luminous emittance of surface** S' (see Fig. 140*b*) is due to quasi-monochromatic sources with identical values of the mean emission frequency, which are located on this surface and are mutually incoherent. A light wave is emitted from each of these sources in the direction of surface S. If the electric field strength of the wave emitted by the source at point P_m' is proportional to $E_m'(t')$, the wave amplitude at point P_1 on surface S will be given by

$$E_{1m}(P_1, t) = E_m'(t - r_{1m}/v)/r_{1m}, \tag{30.78}$$

where v is the velocity of propagation of the wave in the medium between points P'_m and P_{1m}. The factor $t - r_{1m}/v$ in the argument of E'_m takes into account the lag in the oscillations at point P_1 as compared to point P'_m, while r_{1m} in the denominator accounts for the decrease in the amplitude of the electric field strength during wave propagation. The normalization constant has not been included in (30.78). The subscript 1 in E_{m1} stands for point P_{1m} in (30.78).

For quasi-monochromatic emitters, we have

$$E'_m(t') = E'_{0m}(t')\,\mathrm{e}^{-i\omega t'}. \tag{30.79}$$

In this case, Eq. (30.78) assumes the form

$$E_{1m}(t) = E'_{0m}(t - r_{1m}/v)\exp(-i\omega(t - r_{1m}/v)r_{1m}^{-1}. \tag{30.80}$$

The electric field strength at point P_2 is given by a formula similar to (30.80):

$$E_{2m}(t) = E'_{0m}(t - r_{2m}/v)\exp(-i\omega(t - r_{2m}/v)r_{2m}^{-2}. \tag{30.81}$$

At the instant of time t, radiation from all the sources on luminous surface S' arrives at points P_1 and P_2. Hence the resultant electric field strength at these points is given by

$$E_1(t) = \sum_m E_{1m}(t) = \sum_m E'_{0m}(t - r_{1m}/v)\exp(-i\omega(t - r_{1m}/v))r_{1m}^{-1}, \tag{30.82}$$

$$E_2(t) = \sum_{m'} E_{2m'}(t) = \sum_{m'} E'_{0m'}(t - r_{2m'}/v) \times \exp(-i\omega(t - r_{2m'}/v))r_{2m'}^{-1}, \tag{30.83}$$

where the summation is extended to all the emitters on the area element S'.

In complete analogy with (30.5), a suitable change in the reference point for counting time leads to the following expression for the mutual correlation function $\Gamma_{12}(\tau)$ for the quantities E_1 and E_2:

$$\Gamma_{12}(\tau) = \langle E_1(t')E_2^*(t')\rangle$$

$$= \sum_{m,m'} \langle E'_{0m}(t - r_{1m}/v)E'^*_{2m'}(t - r_{2m'}/v)r_{1m}^{-1}r_{2m'}^{-1}\,\mathrm{e}^{i\omega\tau_{mm'}} \tag{30.84}$$

where $\tau_{mm'} = (r_{1m'} - r_{2m'})/v$. Since the point sources of radiation on S' are statistically independent of each other, we obtain

$$\langle E'_{0m}(t - r_{1m}/v)E'^*_{0m'}(t - r_{2m'}/v)\rangle$$

$$= \langle E'_{0m}(t - r_{1m}/v)E'^*_{0m'}(t - r_{2m'}/v)\rangle\,\delta_{mm'}, \tag{30.85}$$

and formula (30.84) assumes the form

$$\Gamma_{12} = \sum_m \langle E'_{0m}(t - r_{1m}/v)E'^*_{0m}(t - r_{2m}/v)\rangle r_{1m}^{-1}r_{2m}^{-1}\,\mathrm{e}^{i\omega\tau_m}, \tag{30.86}$$

where $\omega\tau_m = \omega(r_{1m} - r_{2m})/v = k(r_{1m} - r_{2m})$, $k = \omega/v = 2\pi/\lambda$ is the wave number.

The emission and observation processes considered here are assumed to take place under stationary conditions. Hence we can change the reference point for time on the right-hand side of (30.86) under the summation sign, and write [see (30.5)]

$$\langle E'_{0m}(t - r_{1m}/v)\, E'^{*}_{0m}(t - r_{2m}/v)\rangle = \langle E'_{0m}(t)\, E'^{*}_{0m}(t + \tau_m)\rangle .$$

It is assumed that the coherence time for radiation from all point sources is larger than the time difference for the paths between a source and the points of observation, i.e. $\tau_c > |\tau_m|$. Evaluating (30.86), we can put $\tau_m = 0$ to a fairly high degree of accuracy under this condition and assume that

$$\langle E'_{0m}(t)\, E'^{*}_{0m}(t + \tau_m)\rangle = \langle E'_{0m}(t)\, E'^{*}_{0m}(t)\rangle = 2I'_m, \tag{30.87}$$

where I'_m is the emission intensity corresponding to amplitude E'_{0m} of the electric field strength. Taking (30.87) into account and going over from summation to integration in (30.86), we obtain

$$\Gamma_{12} = 2Q \int_{S'} \frac{I'(P')\, e^{ik(r_1 - r_2)}}{r_1 r_2}\, dS', \tag{30.88}$$

where dS' is an area element on the surface S' at point P' and r_1 and r_2 denote the distance between P' and points P_1 and P_2 respectively. Normalization constant Q, which appears as a result of a transition from summation to integration, need not be determined exactly since it will cancel out in subsequent computations with constants appearing in the denominator.

From (30.88), we obtain the following expressions for self-coherence function (30.6) of light oscillations at points P_1 and P_2:

$$\Gamma_{11} = 2Q \int_{S'} \frac{I'(P')}{r_1^2}\, dS' = 2QI(P_1), \tag{30.89}$$

$$\Gamma_{22} = 2Q \int_{S'} \frac{I'(P')}{r_2^2}\, dS' = 2QI(P_2). \tag{30.90}$$

These are the expressions for the intensity of radiation at points P_1 and P_2 when there is no interference between the radiation emitted by different sources on S'.

Taking into account expressions (30.88), (30.89), (30.90) as well as the relations

$$I(P_1) = I_1 = \int\limits_{S'} \frac{I'(P')}{r_1^2} \mathrm{d}S', \tag{30.91}$$

$$I(P_2) = I_2 = \int\limits_{S'} \frac{I'(P')}{r_2^2} \mathrm{d}S' \tag{30.92}$$

we can transform formula (30.10) for normalized coherence function as follows:

$$\gamma_{12} = \frac{\Gamma_{12}}{\sqrt{\Gamma_{11}\Gamma_{22}}} = \frac{1}{\sqrt{I_1 I_2}} \int\limits_{S'} \frac{I'(P')\,\mathrm{e}^{ik(r_1 - r_2)}}{r_1 r_2} \mathrm{d}S'. \tag{30.93}$$

This relation allows us to evaluate the normalized coherence function at various points in the wave field (if the distribution of the radiation intensity on the surface of an extended quasi-monochromatic source is known) under the assumption that the radiation emitted by the source satisfies the conditions enumerated above. Formula (30.93) expresses the Van-Zittert-Zernike theorem.

Example 30.1. Find the mutual coherence function $\Gamma_{12}(\tau)$ of the wave field for random light radiation with the Lorentz shape of the line. The light flux is composed of a set of plane waves propagating in the positive direction of the Z-axis.

In order to simplify calculations, it is convenient to assume that the radiation is confined to a cavity of quite large size (much larger than the coherence length in the direction of the Z-axis). The boundaries of the cavity can be assumed to be transparent. The electric field strength at a point z inside the cavity at the instant t can be presented as the sum of normal modes of emission:

$$E(z, t) = \sum_k E_k \exp(-i(\omega_k t - kz)), \tag{30.94}$$

where $\omega_k = ck$. Amplitudes E_k are complex quantities whose magnitudes and phases can be determined only statistically.

By definition, the mutual coherence function of the wave field is given by

$$\begin{aligned} \Gamma_{12}(\tau) &= \langle E(z_1, t_1) E^*(z_2, t_2) \rangle \\ &= \sum_{k,k'} \langle E_k E_{k'}^* \rangle \exp[-i(\omega_k t_1 - \omega_{k'} t_2 - kz_1 + kz_2)]. \end{aligned} \tag{30.95}$$

Since for $k \neq k'$ amplitudes E_k and $E_{k'}$ are independent random quantities with zero mean values, we can write

$$\langle E_k E_{k'}^* \rangle = \langle |E_k|^2 \rangle \delta_{kk'}. \tag{30.96}$$

In view of this, Eq. (30.95) assumes the form

$$\Gamma_{12}(\tau) = \sum_k \langle |E_k|^2 \rangle \exp(-i\omega_k \tau), \tag{30.97}$$

where $\tau = t_1 - t_2 - (z_1 - z_2)/c$.

In accordance with (9.28), we can write for a line of the Lorentz shape

$$\langle |E_k|^2 \rangle = A \frac{1}{(\omega_0 - \omega_k)^2 + (\gamma/2)^2}, \tag{30.98}$$

where A is a constant determined by the average energy flux

$$\langle S \rangle = (1/2)\sqrt{\varepsilon_0/\mu_0}\langle |E(z,t)|^2 \rangle = (1/2)\sqrt{\varepsilon_0/\mu_0}\sum_{kk'} \langle E_k E_{k'}^* \rangle$$
$$\times \exp[(\omega_k - \omega_{k'})t - (k - k')z] = (1/2)\sqrt{\varepsilon_0/\mu_0}\sum_k \langle |E_k|^2 \rangle. \tag{30.99}$$

Here we have taken into consideration relation (39.96). For a long cavity ($L \to \infty$), we can go over, as usual, from summation to integration in (30.99):

$$\sum_k \to \frac{L}{\pi}\int_0^\infty dk \to \frac{L}{\pi c}\int_0^\infty d\omega_k. \tag{30.100}$$

Taking (30.98) into account, we obtain for such a transition (when $\gamma \ll \omega_0$)

$$\langle S \rangle = \frac{1}{2}\sqrt{\frac{\varepsilon_0}{\mu_0}}\frac{AL}{\pi c}\int_0^\infty \frac{d\omega_k}{(\omega_0 - \omega_k)^2 + (\gamma/2)^2}$$
$$= \frac{1}{2}\sqrt{\frac{\varepsilon_0}{\mu_0}}\frac{2AL}{\gamma c}. \tag{30.101}$$

Substituting A from (30.101) into (30.98), we obtain

$$\langle |E_k|^2 \rangle = \frac{\gamma c}{L}\sqrt{\frac{\mu_0}{\varepsilon_0}}\langle S \rangle \frac{1}{(\omega_0 - \omega_k)^2 + (\gamma/2)^2}.$$

In this case, (30.97) assumes the form

$$\Gamma_{12}(\tau) = \frac{\gamma c}{L}\sqrt{\frac{\mu_0}{\varepsilon_0}}\langle S \rangle \sum_k \frac{\exp(-i\omega_k \tau)}{(\omega_0 - \omega_k)^2 + (\gamma/2)^2}. \tag{30.102}$$

Going over from summation to integration in this expression, we obtain in accordance with (30.100)

$$\Gamma_{12}(\tau) = \frac{\gamma}{\pi}\sqrt{\frac{\mu_0}{\varepsilon_0}}\langle S\rangle \int_{0}^{\infty} \frac{\exp(-i\omega_k\tau)\,d\omega_k}{(\omega_0-\omega_k)^2+(\gamma/2)^2}.$$

With the help of the residue theory, we obtain the value of the integral:

$$\int_{-\infty}^{\infty} \frac{\exp(-i\omega_k\tau)\,d\omega_k}{(\omega_0-\omega_k)^2+(\gamma/2)^2} = \frac{2\pi}{\gamma}\exp\left(-i\omega_0\tau-\frac{\gamma|\tau|}{2}\right).$$

Hence

$$\Gamma_{12}(\tau) = 2\sqrt{\mu_0/\varepsilon_0}\,\langle S\rangle \exp(-i\omega_0\tau-\gamma|\tau|/2). \qquad (30.103)$$

Thus $\Gamma_{12}(\tau)$ decreases exponentially with increasing line width, as well as with increasing $|\tau|$.

PROBLEMS

5.1. The waves emitted by two point sources separated by a distance $\lambda/2$ have a phase difference π and each of them has an intensity I_0. Find the intensity distribution in the interference pattern.

5.2. The waves emitted by two point sources separated by a distance $\lambda/4$ have a phase difference $\pi/2$ and an intensity I_0 each. Find the intensity distribution in the interference pattern.

5.3. The operating principle of a radio-astronomical interferometer is identical to that of a diffraction grating, the roles of the slits being played by radiowave receivers which produce interference patterns as sources of secondary waves. The linear chain of a radio-astronomical interferometer consists of $N = 40$ receivers separated from each other by a distance $d = 10$ m and operating at a wavelength $\lambda = 21$ cm. Find the angular width of the central peak and the angular distance between the central peak and the principal peaks of the first order.

5.4. Newton's rings are observed under illumination by the *D*-line of sodium with $\lambda = 589$ nm. As the lens is slowly removed from the surface on which it rested at the beginning, the Newton's rings as if contract towards the centre and dark and bright spots appear alternately at the centre. The number of rings being contracted towards the centre can be counted. As the number of such rings approached 500, the visibility of the diffraction pattern deteriorated, and the diffraction pattern completely disappeared when the number of contracted rings was equal to 500. A subsequent increase in the distance between the lens and the surface led to an improved visibility which attained its peak value for 1000 rings. The next minimum and maximum were observed at 1500 and 2000 rings respectively. Such a behaviour of the interference pattern is due to the doublet nature of the *D*-line. Find the separation between the lines of this doublet.

5.5. The optical path length in cell Q of a Mach-Zehnder interferometer (see Fig. 108) is $l = 20.4$ cm. The cell is almost completely evacuated.

A dark spot is observed at the centre of the interference pattern. The wavelength of the light emitted by a monochromatic source is $\lambda = 589$ nm. Air is slowly let into the cell through a valve. As a result, the pressure in the cell increases, the radii of rings in the interference pattern change and the illuminance at the centre of the interference pattern increases and decreases alternately. The number of brightenings is counted by the experimenter. By the time the pressure of air in the cell reached the value 1 atm, 101 brightenings had been counted. A dark spot was observed at the centre of the interference pattern obtained at a pressure of 1 atm. Find the refractive index of air at atmospheric pressure.

5.6. Find the lateral displacement of the central peak of the interference pattern formed by rays emerging from two slits if one of the slits is covered by a plane-parallel plate with thickness q and the refractive index of its material n. The distance between the slits and the screen is l (see Fig. 113).

5.7. A plane-parallel plate rests on a horizontal glass table. One edge of the plate is in contact with the surface of the table, while the other is slightly raised to form a very small angle α between the table surface and the plate surface. A monochromatic wave of wavelength 589 nm is normally incident on the plate and the table from above. The distance between fringes of equal thickness is 5 mm. Find the angle α.

5.8. In the Twyman-Green interferometer (see Fig. 110), converging lens L is removed and the reflection takes place directly from the convex mirror having a radius of curvature 5 m. The wavelength of the light emitted by the monochromatic source is 550 nm. The interference pattern is observed on the screen by naked eye without the help of any lens. The distances from the centre of the dividing plate to A_1 and to the convex mirror A_2 are equal, and hence a bright spot is observed at the centre of the interference pattern. Find the radius of the ninth bright ring.

5.9. Find the radius of the fourth ring in the interference pattern formed by the Fabry-Perot interferometer in the focal plane of a lens with focal length $f = 0.5$ m. Distance d between the interferometer plates is 1 cm, and the wavelength of the monochromatic radiation is $\lambda = 693$ nm.

5.10. A diameter is passed through the centre of a thin lens with focal length 10 cm. Parallel to this diameter, a 1 mm thick plane strip is cut from the lens symmetrically with respect to the diameter. The remaining parts of the lens are tightly pressed together. Find the wavelength of light from a point source located at a distance of 8 cm from the lens if the distance between adjacent maxima in the interference pattern formed on a screen at a distance of 160 cm from the lens is 0.3 mm.

5.11. For what distance between the mirrors in the Fabry-Perot interferometer can the doublet of sodium D-line be separated ($\lambda = 589$ nm, $\Delta\lambda = 0.6$ nm)?

5.12. Rings of equal inclination are observed in the Michelson interferometer from a mercury lamp emitting light of wavelength $\lambda = 435.8$ nm. The interference pattern is blurred when the mobile mirror is displaced by 47.5 mm. Find the linewidth.

5.13. Find the number of rulings per millimetre of a diffraction grating such that the sodium doublet lines in the second order are separated by a distance of 1 mm on a photographic plate obtained by using a camera with focal length 100 cm ($\Delta\lambda = 0.6$ nm and $\lambda = 589$ nm for the sodium doublet).

5.14. A soap film of thickness much smaller than a wavelength is formed in air. Will it appear bright or dark in reflected light?

5.15. The rays emerging from a Lummer-Gehrcke plate can be treated as practically grazing along the surface. Determine the dispersion region for the plate taking into account the dispersion of light. The refractive index of the material of the plate is n, the plate thickness is h, and the wavelength of light is λ.

5.16. Diffused monochromatic light with $\lambda = 0.5$ μm is incident on a thin film with refractive index $n = 1.5$. Find the film thickness if the rays corresponding to neighbouring maxima near the angle of reflection of 45° form an angle of 2°.

5.17. The interference fringes in the Jamin interferometer are observed in the focal plane of an objective with focal length 15 cm. The refractive index of the material of the plates is $n = 1.5$ and the plates are 2 cm thick. The angle between the plates is 1′. Light is incident on the plates at an angle of 45°. The interference fringes are separated by a distance of 5 mm. Find the wavelength of light producing the interference pattern.

5.18. A split lens is prepared from a lens with focal length $F = 15$ cm, and its halves are separated $a = 1$ mm apart. Find the separation between interference fringes if the light source is a slit located at a distance of $L = 30$ cm from the split lens while the screen is at a distance of $l = 5$ m from the lens. The wavelength of light is $\lambda = 0.5$ μm.

5.19. What is the minimum distance between the slits in Young's experiment so that the interference pattern formed by the solar radiation can be observed? The angular dimension of the Sun is 32′, and the wavelength of the radiation is assumed to be 0.55 μm [see (27.25)].

5.20. A Fresnel biprism with a small angle of refraction α is placed at a distance L from the light source and of a distance l from the screen. Find the number of interference fringes if the refractive index of the material of the biprism is n and the wavelength is λ.

5.21. Newton's rings are formed between two planoconvex lenses whose convex faces are in contact. The radii of curvature of these surfaces are R_1 and R_2. Find the radius of the m-th dark ring in reflected light. The wavelength of the light is λ.

5.22. During measurements of the angular dimensions of an astronomic body with the help of a stellar interferometer it was found that the visibility of the interference pattern attains successive minima as the distance between the mirrors is increased by 15 cm. Measurements are being made at a wavelength of 550 μm. Find the angular dimensions of the body.

5.23. Interference of N rays is observed in a Lummer-Gehrcke plate. The refractive index of the glass is 1.71 and the wavelength is $\lambda = 540$ nm. Find the observed order of interference and the resolving power of the plate under these conditions.

5.24. While observing the interference from solar radiation at wavelength $\lambda = 0.6$ μm with a varying distance between the slits in Young's experiment, it was found that the visibility is at its lowest value when the distance between the slits is 0.064 mm. What is the angular dimension of the Sun?

5.25. A point source (slit) illuminating a Lloyd's mirror is placed at a height of 1 cm above its plane and at a distance of 5.5 m from the screen which is perpendicular to the plane of the mirror and is in close contact with the latter on the side opposite to the light source. The mirror is 50 cm long. Find the distance between maxima in the interference pattern and the number of bright and dark fringes. The wavelength of light is 0.5 μm.

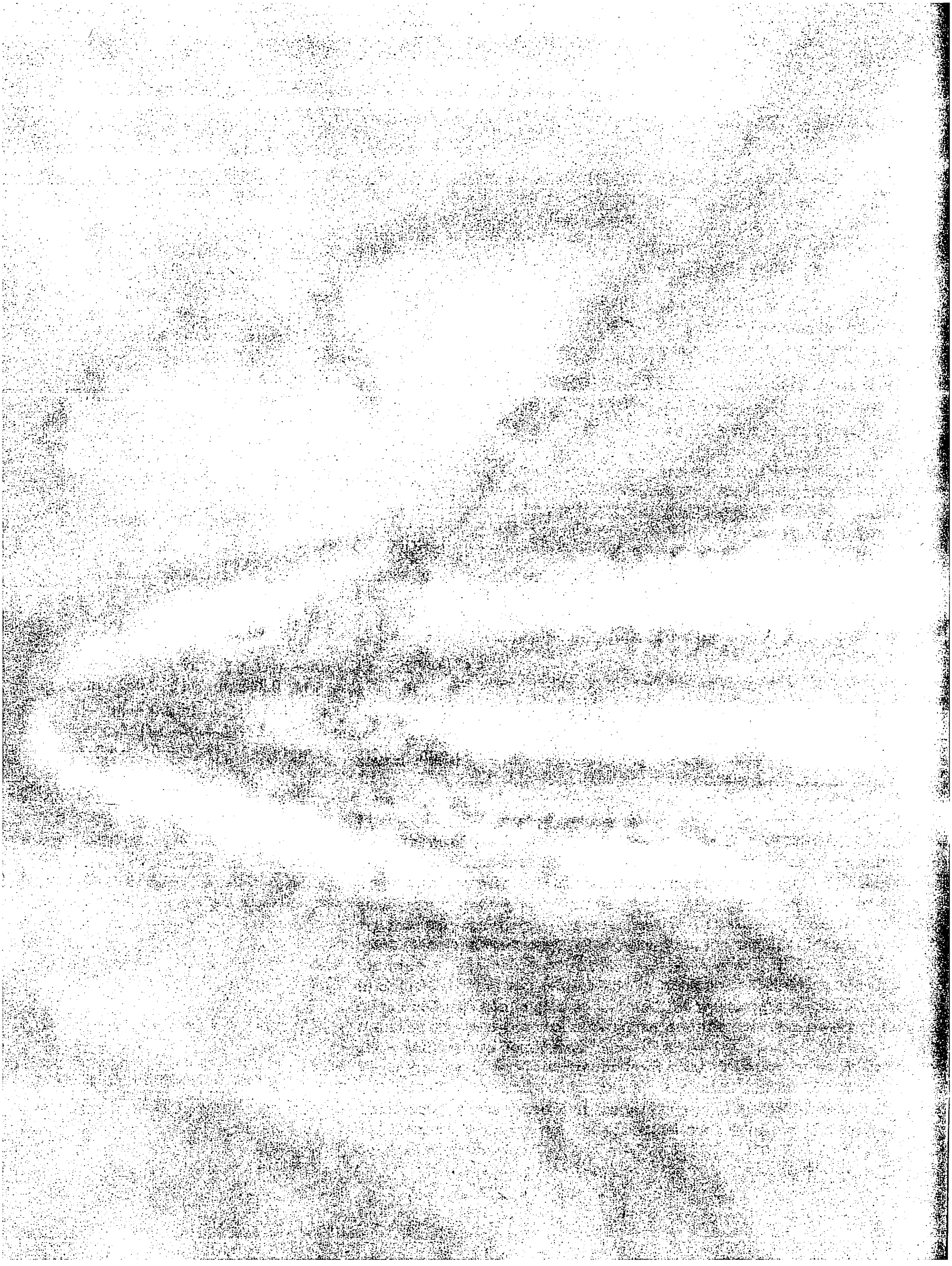
5.26. The acute angle at the apex of a Fresnel biprism is 20′, and the thickness of the biprism can be neglected in calculations. A point source is located at the symmetry axis of the biprism perpendicular to the largest side and at a distance of 10 cm from the biprism, while the screen is perpendicular to this axis and is at a distance of 2.9 m from the biprism. The wavelength of light is 0.5 μm. How many bright and dark fringes are formed on the screen?

5.27. A Billet split lens with a focal length of 20 cm has a gap of 0.5 mm between the lens halves. The distance between the split lens and a point source (a horizontal slit) is 40 cm, while the distance between the lens and the screen is 60 cm. Find the distance between the maxima in the interference pattern and the number of bright fringes on the screen (the diameter of the split lens is quite large, say, larger than 1 cm). The wavelength of the light is $\lambda = 0.5$ μm.

5.28. Find the coherence function for a random light with the Gaussian frequency distribution described by formula (10.17).

ANSWERS

5.1. $4I_0\sin^2\left(\frac{\pi}{2}\sin\theta\right)$. 5.2. $2I_0\left[1+\sin\left(\frac{\pi}{2}\sin\theta\right)\right]$. 5.3. 3′36″; 1°12′. 5.4. $\Delta\lambda = 0.6$ nm. 5.5. 1.000293. 5.6. $lq(n-1)/a$. 5.7. 0.59×10^{-4} rad. 5.8. 5 mm. 5.9. 8.3 mm. 5.10. 0.6 μm, 5.11. 0.29 mm. 5.12. 0.002 nm. 5.13. 600. 5.14. Dark. 5.15. $[\lambda^2\sqrt{n^2}-1/(2h)][\lambda n(\mathrm{d}n/\mathrm{d}\lambda)-(n^2-1)]^{-1}$. 5.16. 18.7 μm. 5.17. 434 nm. 5.18. $\lambda[l(L-f)-Lf]/(La) =$ 1.8 mm. 5.19. 0.03 mm. 5.20. $4Ll\alpha^2(n-1)^2/[\lambda(L+l)]$. 5.21. $\sqrt{m\lambda/(1/R_1+1/R_2)}$. 5.22. 3.7×10^{-6} rad. 5.23. 1.03×10^5, 6.68×10^6. 5.24. 32′. 5.25. 0.14 μm; 7 dark, 7 bright and 1 dark central. 5.26. 7 bright, 6 dark. 5.27. 0.1 mm; 13. 5.28. $\exp(-\sigma^2\tau^2/2)$.



Chapter 6

Diffraction

Basic idea:

Interference of radiation emitted by secondary sources on the wave front leads to diffraction.

SEC. 31. FRESNEL ZONE METHOD

The basic ideas behind the analysis of secondary sources are described by considering the example of Fresnel zones for various situations.

HUYGENS-FRESNEL PRINCIPLE (1818). The hypothesis that each point on a wave front is a source of secondary waves [Huygens'principle, see (8.27)] was supplemented by Fresnel with the statement that these secondary sources are mutually coherent and the waves emitted by them interfere (Fig. 142). Thus, while analyzing the propagation of waves we must take into consideration their amplitudes and phases. This allows us to study the problem of light intensity Fresnel was sure that the amplitude of the secondary wave depends on the angle between the normal to the primary wave front and the direction towards a point on the secondary wave front. The amplitude has its peak value in the direction of the normal, while in the perpendicular direction, i.e. along the tangent to the primary wave front, the amplitude is equal to zero. A more precise form of this dependence was not known at that time.

FRESNEL ZONES. A typical example of propagation of finite light beams is shown in Fig. 143. A spherical (or plane) wave is incident on an opaque screen with an aperture in the middle. We have to find the distribution of light intensity behind the screen. To solve this problem with the help of the Huygens-Fresnel principle, we make two assumptions: (1) the opaque parts of the screen are not sources of secondary waves; (2) the points of the wave front in the aperture are sources of the secondary waves in the same way as if the opaque screen were missing.

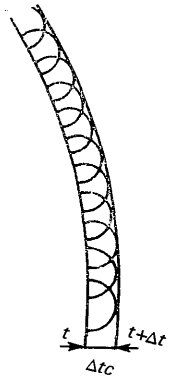


Fig. 142. Illustrating Huygens' principle.

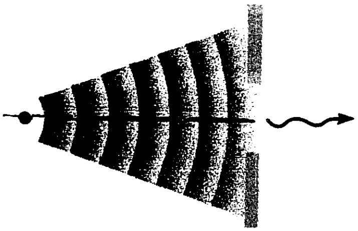


Fig. 143. Propagation of light through a finite-size aperture.

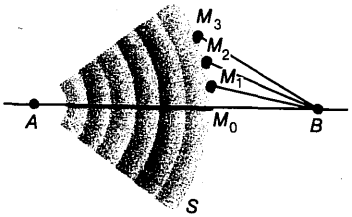


Fig. 144. To the determination of Fresnel zones.

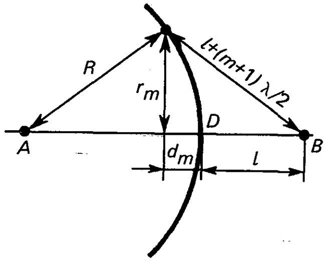


Fig. 145. To the calculation of radii and areas of Fresnel zones.

Let A be a source of a spherical wave and S be the wave front at a certain instant of time (Fig. 144). Let us find the wave intensity at point B with the help of the Huygens-Fresnel principle. We divide surface M into annular zones in such a way that the distances between zone boundaries and B differ by $\lambda/2$. Denoting the zone boundaries by $M_0, M_1, M_2, \ldots,$ we can write this condition in the form

$$\begin{aligned} &M_1B - M_0B = \lambda/2, \\ &M_2B - M_1B = \lambda/2, \\ &\ldots\ldots\ldots\ldots\ldots \qquad (31.1) \\ &M_nB - M_{n-1}B = \lambda/2. \end{aligned}$$

The radii of the zones are determined with the help of a construction shown in Fig. 145. Here $r_1, r_2, \ldots, r_m$ are the zone radii, R is the radius of curvature of the spherical wave front, D is the point of intersection of the wave front with straight line AB, and $d_1, d_2, \ldots, d_m$ are the distances between D and the projections of the respective zone boundaries onto straight line AB. **The central zone is called the zeroth zone.** For radius r_m, we write the condition

$$r_m^2 = R^2 - (R - d_m)^2 = [l + (m+1)\lambda/2]^2 - (l + d_m)^2, \qquad (31.2)$$

whence we obtain

$$\begin{aligned} d_m &= \frac{l(m+1)}{R+l}\frac{\lambda}{2}, \\ r_m^2 &= \frac{Rl(m+1)\lambda}{R+l}, \end{aligned} \qquad (31.3)$$

with an error not exceeding the quantity λ^2. Consequently, the area of the zeroth zone is given by

$$S_0 = \pi r_0^2 = \pi Rl\lambda/(R+l). \qquad (31.4)$$

For the total area of the zeroth and first zone, we obtain

$$S_{(0+1)} = \pi R/2\lambda/(R+l). \qquad (31.5)$$

Hence the area of the first zone is

$$S_1 = S_{(0+1)} - S_0 = \pi Rl/(R+l), \qquad (31.6)$$

i.e. is equal to the area of the zeroth zone. The areas of all other zones are also given by the same expression [see (31.3)]. In these calculations, the curvature of the wave front surface is neglected, and it is assumed that the area of the annular zone on the spherical surface of the wave front is equal to the area of its projection on a plane perpendicular to straight line AB. This does not introduce any significant errors in the calculations if the radii of Fresnel zones are much smaller than the radii of curvature of the wave front. Considering the smallness of the wavelengths, we conclude from (31.3) that these conditions are satisfied quite well for a very large number of Fresnel zones.

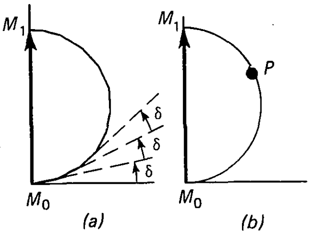


Fig. 146. Addition of complex amplitudes in zeroth Fresnel zone: when the zone is conditionally divided into finite regions (*a*); when the phase changes continuously (*b*).

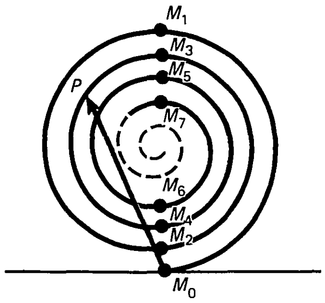


Fig. 147. Spiral.

GRAPHIC COMPUTATION OF AMPLITUDE. In order to obtain quantitative results, we must carry out a more exact analysis of the interference of secondary waves at point B.

We divide each zone into a large number n of subzones. The phase of the waves changes by π from the beginning to the end of a zone, hence the phase changes from one subzone to the next by $\delta = \pi/n$. We denote by E_0 the amplitude of the wave arriving at B from each subzone. The phase of the wave arriving at B from the zeroth zone (beginning from M_0) is assumed to be zero. Then, taking into account the interference, the complex amplitude of the wave arriving at point B from the zeroth zone is given by

$$E = E_0 + E_0 e^{i\delta} + E_0 e^{i2\delta} + \dots \qquad (31.7)$$

The analytic sum of the amplitudes in (31.7) may be obtained graphically (Fig. 146*a*). As the number of subzones increases to infinity ($n \to \infty$), the broken curve changes into a smooth one (Fig. 146*b*). The length of segment M_0P is proportional to the wave amplitude at point B (see Fig. 145) if the part of the zeroth zone from the centre to the boundary corresponding to point P is open. The length $|M_0M_1|$ is proportional to the amplitude when the zeroth zone is completely open. The graphic construction of the amplitudes taking into account the subsequent zones is carried out in a similar manner. The only factor that must be taken into account is that the value of E_0 is slightly reduced as we move away from the point (i.e. as the phase in the exponential multiplier increases). Consequently, the continuous curve does not close but assumes a helical shape (Fig. 147). This allows us to determine the amplitude when any number of zones and their parts are open. For example, segment $|M_0P|$ in Fig. 147 is proportional to the amplitude when the zeroth, first, second, and a part of the third zone are open. The lengths $|M_0M_1|$, $|M_0M_2|$, ... are proportional to the amplitudes when the zeroth, zeroth and first, etc. zones are open. When all the zones are open, the amplitude is proportional to $|M_0M_\infty| \approx |M_0M_1|/2$. This shows that when only the zeroth zone is open, the wave amplitude at point B is about twice and the intensity is about four times the corresponding values when all zones are open. When the zeroth and the first zones are open, the amplitude is proportional to $|M_0M_2|$, and hence the intensity is very low. If the radius of the aperture on which the wave is incident increases continuously, the wave intensity at point B also changes continuously. With the help of the spiral shown in Fig. 147, we can plot the change in wave amplitude as a function of r (Fig. 148).

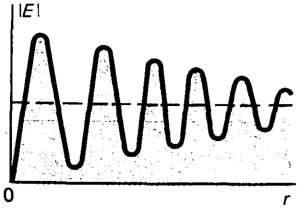


Fig. 148. Dependence of amplitude on aperture radius.

POISSON'S SPOT. If an opaque circular screen is placed in the path of a light wave (Fig. 149), a bright spot is observed in the shadow of the screen. This spot is called Poisson's spot. The emergence of the bright spot follows directly from an analysis by the Fresnel zone technique. The screen blocks a certain number of Fresnel zones starting from the zeroth zone. However, the zones beyond those covered by the screen produce an illumination at point B whose intensity can be calculated from the spiral shown in Fig. 147. For example, let us suppose that a certain number of zones has been blocked by the screen. If the screen is not very small, the number of such zones will be very large. Hence the point corresponding to the screen radius will lie on the spiral (Fig. 147) not far from M_∞ (point P). Obviously, the wave amplitude at point B is proportional to $|PM_\infty|$. Thus, the wave as if bends around the opaque screen. The phenomenon of bending of waves round obstructions is called diffraction.

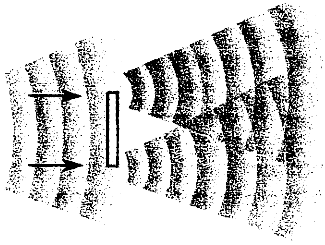


Fig. 149. Emergence of Poisson's spot.

For large-sized opaque screens, the intensity of Poisson's spot is quite low, since point P (see Fig. 147) is quite close to M_∞. Moreover, the light must be highly coherent since otherwise there would be no interference between rays from different parts of the zones. To observe the diffraction pattern, quite small screens must be used. With the help of the laser radiation, however, diffraction can be clearly demonstrated on large screens by using comparatively large obstructions.

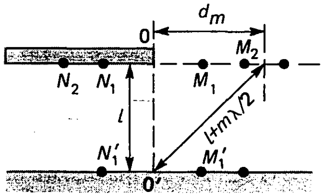


Fig. 150. Fresnel zone calculation for diffraction at the knife-edge of a semi-infinite opaque screen.

DIFFRACTION AT THE KNIFE-EDGE OF A SEMI-INFINITE SCREEN. Let us consider the incidence of a plane wave on a semi-infinite screen (Fig. 150). The zones are counted from the edge of the screen, and the closest zone is treated as the first one. We denote by d_m the distance from the origin to the far end of the m-th zone. It can be seen from Fig. 150 that with an error not exceeding quantities of the order of λ^2, we have

$$d_m^2 = (l + m\lambda/2)^2 - l^2 \approx lm\lambda, \quad d_m = \sqrt{lm\lambda}. \qquad (31.8)$$

Fresnel zones are fringes with their boundaries parallel to the edge of the screen. The areas of the zones are in proportion to their widths, i.e. are related through the ratio

$$d_1 : (d_2 - d_1) : (d_3 - d_2) : \ldots = 1 : (\sqrt{2} - 1) : (\sqrt{3} - 2) : \ldots \qquad (31.9)$$

Since $(\sqrt{2} - 1) = 0.41$ and $(\sqrt{3} - \sqrt{2}) = 0.32$, it is obvious that the areas of Fresnel zones decrease rapidly as we move

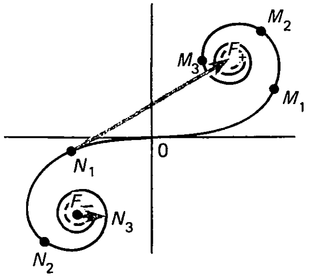


Fig. 151. Cornu's spiral.

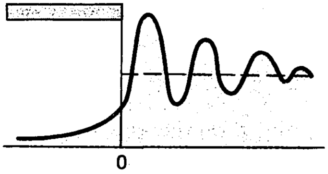


Fig. 152. Distribution of illumination intensity behind a semi-infinite opaque screen.

away from the screen edge. This means that as a result of vector summation of amplitudes in accordance with (31.7), the values of E_0 decrease with increasing phase much more rapidly than in the case of a circular aperture. Hence the graphic method of addition of amplitudes leads to a helix different from the one shown in Fig. 147. This spiral is called Cornu's spiral.

The lengths $|OM_1|, |OM_2|, \ldots$ of line segments in Fig. 151 are proportional to illumination intensity at point O if the zones located at $|OM_1|, |OM_2|, \ldots$ in Fig. 150 are open. If the entire semi-infinite space to the right of O is open, the illumination intensity at point O will be proportional to $|OF_+|$. The lengths $|ON_1|, |ON_2|, \ldots$ of the line segments (Fig. 151) are proportional to the illumination intensity at point O when the zones located at $|ON_1|, |ON_2|, \ldots$ in Fig. 150 are open. If the entire semi-infinite space to the left of O is open, i. e. if the opaque screen is located to the right of point O, the illumination intensity at point O will be proportional to $|OF_-|$. If there is no screen at all, i. e. if the entire space is open, the illumination intensity at O will be proportional to $|F_-F_+|$.

The illumination intensity at other points is determined with the help of Cornu's spiral, assuming that point O on this spiral corresponds to the point under consideration. For example, to find the illumination intensity at point M_1' (see Fig. 150), we assume that point O on Cornu's spiral (Fig. 151) corresponds to point M_1.

Consequently, a certain point on Cornu's spiral, say N_1, to the left of O in Fig. 150 corresponds to the edge of the screen. Hence the illumination intensity is given by the length of the line segment $|N_1F_+|$ (Fig. 151). To determine the illumination intensity at point N_1' (see Fig. 150), we assume that point O on Cornu's spiral corresponds to this point. Consequently, the edge of the screen corresponds to a certain point on Cornu's spiral, say M_1, to the right of O. Therefore, the illumination intensity at N_1' is given by the length of the line segment M_1F_+. This method can be used to determine the illumination intensity at any point not only for an infinite screen covering the entire semi-infinite space, but also for any system of screens in the form of infinitely long opaque strips whose edges are parallel to one another. For this purpose, we must obtain the vector sum of the illumination intensity amplitudes from open regions which transmit a wave. The wave amplitude at each segment is obtained with the help of Cornu's spiral. However, the geometrical construction is found to be quite cumbersome in this case, and hence analytic methods, which are more effective in such situations (see Sec. 34), are preferred, whereas for a semi-infinite screen the geometrical construction with the help of Cornu's spiral is quite simple. Figure 152 shows the plot of variation of amplitude behind a semi-infinite screen.

ZONE PLATE AS A LENS. In accordance with (31.3), the radius of the *m*-th plate is given by

$$r_m = [Rl(m+1)\lambda/(R+l)] \quad (m = 0,\ 1,\ 2,\ \dots). \tag{31.10}$$

We close all odd zones, leaving all even zones open. As a result, we get a device which is called a zone plate (Fig. 153). The amplitude of a light wave transmitted through the plate at point *B* on the axis can be calculated with the help of a spiral shown in Fig. 147. The amplitude due to the zeroth open circular zone is given by vector M_0M_1, the amplitude from the second annular zone is given by vector M_2M_3, from the fourth zone by vector M_4M_5, and so on. All these vectors have the same direction, i. e. the phases of the complex amplitudes differ by $2\pi m$, where *m* is an integer. Hence augmented interference of the waves takes place. As a result, a significant enhancement of the light intensity takes place at point *B* on the axis, i. e. light is focussed at this point. The zone plate thus behaves like a lens.

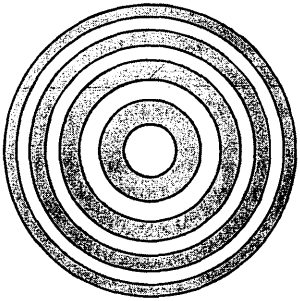


Fig. 153. Zone plate.

We shall assume that the rays incident on the plate are parallel, i. e. $R = \infty$. In this case, the point on the axis where the rays converge behaves as a focus of the lens identically placed as a zone plate, if its focal length is $f = l$. As $R \to \infty$, formula (31.10) assumes the form

$$r_m^2 = (m+1)/\lambda. \tag{31.11}$$

Hence, the focal length is

$$f = l = r_m^2/[(m+1)\lambda]. \tag{31.12}$$

Relation (31.10) can be presented as the formula

$$\frac{1}{R} + \frac{1}{l} = \frac{1}{r_m^2/(m+1)\lambda]} = \frac{1}{f}, \tag{31.13}$$

which shows that relation (23.27) characterizing an object and its image obtained with the help of a lens is valid for a source and a point on the axis behind the zone plate. This means that a zone plate behaves like a converging lens and can be used, for example, to form an image and carry out other operations characteristic of a lens. This serves as an experimental confirmation of the concept of Fresnel zones.

DRAWBACKS OF THE FRESNEL ZONE METHOD. The Fresnel zone method leads to results that are in good agreement with the cases of practical importance, when the size of the obstruction is much more than the wavelength.

!

The physical nature of the Huygens principle is that light is a continuous process in space. The Huygens principle is equivalent to the description of light propagation with the help of differential equations which, however, were not known at that time.
Huygens' idea that each point of a wave front is a source of secondary wave was supplemented by Fresnel's statement that these sources are mutually coherent, and the secondary waves emitted by them interfere.

?

What is the qualitative dependence of the intensity of Poisson's spot on the distance from an opaque screen?
Why is the intensity at the focal point farthest from a zone plate the highest among the intensities at all the focal points of the plate? List the main difficulties encountered in the Fresnel zone method.

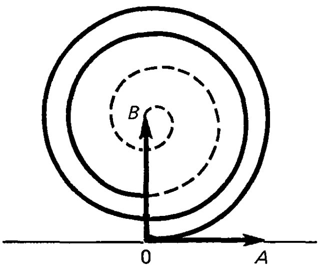


Fig. 154. Difference between the actual phase of a wave and the phase calculated by the Fresnel method.

However, this method suffers from a number of serious drawbacks.

1. It does not describe the damping of amplitude of secondary waves as a function of the direction of propagation. This dependence has to be postulated in order to ensure that no inverse wave is formed.

2. The Fresnel method does not give a correct phase of the wave. By definition, the phase on the wave front is taken equal to zero. Hence the wave amplitude is given by the vector **OA** (Fig. 154). The amplitude calculated in accordance with the Fresnel method is given by vector **OB**, i. e. the phase calculated with the help of Fresnel's method differs from the actual phase of the wave by $\pi/2$. Although this phase difference is insignificant for many phenomena of practical importance which depend on the modulus of the amplitude, it is fundamental from the theoretical point of view and must be explained. This was made possible only by considering a more rigorous diffraction theory based on Kirchhoff's integral.

SEC. 32. KIRCHHOFF'S APPROXIMATION

Formulas of the diffraction theory are derived with the help of Kirchhoff's approximation.

GREEN'S FORMULA. If in the Gauss theorem

$$\int_V \operatorname{div} \mathbf{A} \, \mathrm{d}V = \oint_S \mathbf{A} \cdot \mathrm{d}\mathbf{S} \tag{32.1}$$

we put

$$\mathbf{A} = \Phi \operatorname{grad} G - G \operatorname{grad} \Phi, \tag{32.2}$$

we obtain the relation

$$\int_V (\Phi \nabla^2 G - G \nabla^2 \Phi) \, \mathrm{d}V = \oint_S \left(\Phi \frac{\partial G}{\partial n} - G \frac{\partial \Phi}{\partial n} \right) \mathrm{d}S. \tag{32.3}$$

Here $(\partial G/\partial n)\,\mathrm{d}S = \operatorname{grad} G \cdot \mathrm{d}\mathbf{S}$, $(\partial \Phi/\partial n)\,\mathrm{d}S = \operatorname{grad} \Phi \cdot \mathrm{d}\mathbf{S}$, i. e. $\partial G/\partial n$ and $\partial \Phi/\partial n$ are the derivatives with respect to length parallel to the outward normal to closed surface S, and V is the volume bounded by surface S.

Formula (32.3) is called Green's second formula. It is applicable when the functions Φ and G, as well as their first and second partial derivatives, are continuous in volume V and on surface S.

HELMHOLTZ-KIRCHHOFF THEOREM. The propagation of light waves in free space is described by Eq. (2.12). In the case of a monochromatic wave, we can put

$$\Phi(\mathbf{r}, t) = \Phi(\mathbf{r})\, \mathrm{e}^{-i\omega t}. \tag{32.4}$$

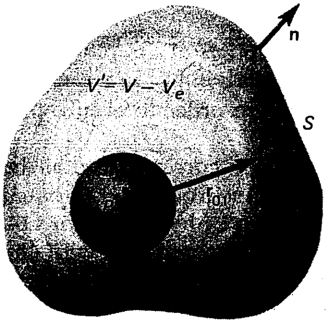


Fig. 155. To the derivation of the Kirchhoff-Helmholtz integral theorem.

This gives the following equation for the amplitude which depends only on spatial coordinates:

$$\Delta^2\Phi + k^2\Phi = 0, \tag{32.5}$$

where $k^2 = \omega^2/c^2 = (2\pi/\lambda)^2$, and λ is the wavelength.

Let us apply formula (32.3) to the volume V bounded by surface S (Fig. 155). We denote a fixed point inside V by P_0, and a moving point (any point in space other than P_0) by P_1. Taking P_0 as the reference point, we can characterize the coordinates of point P_1 by the radius vector $\mathbf{r}_{01}$. A direct verification shows that the function

$$G(P_1) = \exp(ikr_{01})/r_{01} \tag{32.6}$$

satisfies equation (32.5) for all points other than P_0. In notation (32.6) for $G(P_1)$, we have taken into account the fact that point P_0 has been chosen as the origin of coordinates, and there is no need to specify it. When $G(P_1)$ is substituted into (32.5), the derivatives are obtained with respect to the coordinates of point P_1. The function Φ describing an electromagnetic wave is continuous at all points.

Since at point $P_1 = P_0$ the function $G(P_1)$ tends to infinity, while its derivative undergoes a discontinuity, we cannot apply Green's formula for the entire volume to it. Hence we surround point P_0 by a small sphere of radius ε with its centre coinciding with P_0. We denote the surface area of this sphere by S_ε and the volume bounded by it as V_ε. Outside the volume V_ε, the function $G(P_1)$ satisfies all the conditions of applicability of (32.3). Considering that

$$\int\limits_{V-V_\varepsilon} (\Phi\nabla^2 G - G\nabla^2\Phi)\,\mathrm{d}V = \int\limits_{V-V_\varepsilon} (-k^2\Phi G + k^2\Phi G)\,\mathrm{d}V = 0, \tag{32.7}$$

we obtain

$$\int\limits_{S+S_\varepsilon} \left(\Phi\frac{\partial G}{\partial n} - G\frac{\partial\Phi}{\partial n}\right)\mathrm{d}S = 0. \tag{32.8}$$

It should be observed that the outward normal $\mathbf{n}$ to the surface S_ε which restricts the volume $V' = V - V_\varepsilon$ is directed into the volume V_ε (Fig. 155). It follows from (32.7) and (32.3) that

$$-\int\limits_{S_\varepsilon} \left(\Phi\frac{\partial G}{\partial n} - G\frac{\partial\Phi}{\partial n}\right)\mathrm{d}S = \int\limits_{S} \left(\Phi\frac{\partial G}{\partial n} - G\frac{\partial\Phi}{\partial n}\right)\mathrm{d}S. \tag{32.9}$$

For points P_1 on surface S, we can write

$$\frac{\partial r_{01}}{\partial n} = \mathrm{grad}_1\, r_{01}\cdot\mathbf{n} = \cos(\mathbf{n}, \mathbf{r}_{01}). \tag{32.10}$$

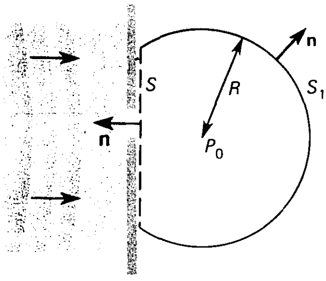


Fig. 156. To the derivation of radiation condition.

Hence

$$\partial G/\partial n = \cos(\mathbf{n}, \mathbf{r}_{01})(ik - 1/r_{01})G. \tag{32.11}$$

The subscript "1" on the gradient operator in Eq. (32.10) indicates that the gradient is taken over coordinate points P_1, i.e. $\text{grad}_1\, r_{01} = \mathbf{r}_{01}/r_{01}$. Obviously, $\text{grad}_0\, r_{01} = \mathbf{r}_{10}/r_{10}$. Hence $\text{grad}_1\, r_{01} = -\text{grad}_0\, r_{01}$. For points P_1 on the surface S_ε, we have

$$G(P_1) = e^{ik\varepsilon}/\varepsilon, \quad \partial G/\partial n = -\partial G/\partial\varepsilon = (1/\varepsilon - ik)G. \tag{32.12}$$

Since Φ and its derivatives are continuous for $\varepsilon \to 0$, we get

$$\int_{S_\varepsilon \to 0} \left(G\frac{\partial\Phi}{\partial n} - \Phi\frac{\partial G}{\partial n}\right)\mathrm{d}S = \left\{\left[\frac{e^{ik\varepsilon}}{\varepsilon}\frac{\partial\Phi}{\partial n} - \Phi\frac{e^{ik\varepsilon}}{\varepsilon}\left(\frac{1}{\varepsilon} - ik\right)\right]4\pi\varepsilon^2\right\}_{\varepsilon\to 0} = -4\pi\Phi(P_0). \tag{32.13}$$

Hence [see (32.9)]

$$\Phi(P_0) = \frac{1}{4\pi}\int_S \left[\frac{e^{ikr_{01}}}{r_{01}}\frac{\partial\Phi}{\partial n} - \Phi\frac{\partial}{\partial n}\left(\frac{e^{ikr_{01}}}{r_{01}}\right)\right]\mathrm{d}S. \tag{32.14}$$

This formula allows us to calculate the values of the function Φ at any point inside a volume if we know the value of the function and of its derivative with respect to the normal to the surface bounding this volume. Relation (32.14) is called the Helmholtz-Kirchhoff integral theorem and forms the basis of the scalar theory of diffraction.

The function Φ in (32.14) must satisfy the wave equation. Hence the values of Φ and $\partial\Phi/\partial n$ on the surface cannot be given arbitrary values. Consequently, (32.14) is not a formula that can be used to calculate $\Phi(P_0)$, but is rather an integral equation in Φ, and the apparent simplicity with which $\Phi(P_0)$ can be evaluated is deceptive.

The physical meaning of the function $G = \exp(ikr_{01})/r_{01}$ appearing in (32.14) is obvious from a comparison of this function with the right-hand side of (2.43). It can be seen that G is the coordinate-dependent part of the spherical wave emanating from point P_1.

RADIATION CONDITION. Let us consider a typical situation in which diffraction is studied (Fig. 156). Light is incident on a large (infinite) screen S with one or several apertures, and the intensity distribution of light passing through the apertures, i.e. the diffraction of light in the space behind the screen, is observed. The value of $\Phi(P_0)$ can be expressed through formula (32.14), and the surface confining the volume consists of two parts: a plane surface S and a part

S_1 of the surface of a sphere having a radius R and its centre at the point P_0. Integral (32.14) is presented as the sum of two integrals, one over the surface S and the other over S_1. Let us evaluate the integral over S_1 as $R \to \infty$. Since

$$\frac{\partial}{\partial n}\left(\frac{e^{ikr_{01}}}{r_{01}}\right) = \frac{\partial}{\partial R}\left(\frac{e^{ikR}}{R}\right) = \left(ik - \frac{1}{R}\right)\frac{e^{ikR}}{R}, \tag{32.15}$$

for points on the surface S_1, we have

$$Q = \int_{S_1}\left[\frac{e^{ikr_{01}}}{r_{01}}\frac{\partial\Phi}{\partial n} - \Phi\frac{\partial}{\partial n}\left(\frac{e^{ikr_{01}}}{r_{01}}\right)\right]dS$$
$$= \frac{e^{ikR}}{R}\int_{S_1}\left[\frac{\partial\Phi}{\partial n} - \Phi\left(ik - \frac{1}{R}\right)\right]dS. \tag{32.16}$$

We denote by Ω the solid angle at which the surface S_1 is seen from point P_0. Considering that $dS = R^2 d\Omega$ in (32.16), we obtain

$$Q = e^{ikR}\int_{S_1} R\left[\frac{\partial\Phi}{\partial n} - \left(ik - \frac{1}{R}\right)\Phi\right]d\Omega. \tag{32.17}$$

This integral is equal to zero as $R \to \infty$, if the radiation condition is satisfied:

$$\lim_{R\to\infty} R\left[\frac{\partial\Phi}{\partial n} - \left(ik - \frac{1}{R}\right)\Phi\right] = 0. \tag{32.18}$$

In this case, formula (32.14) does not change its form for the situation presented in Fig. 156. However, S now stands not for the entire closed surface enveloping point P_0, but rather a plane infinite surface which is also marked in Fig. 156 by S.

We can present certain arguments showing that condition (32.18) is indeed satisfied under real physical situations. For a finite-sized aperture and a finite distance between the aperture and point P_0, the function Φ must describe a spherical wave for $R \to \infty$, i.e. must have the form

$$\Phi(R) \propto e^{ikR}/R. \tag{32.19}$$

Since

$$\frac{\partial\Phi}{\partial n} = \frac{\partial\Phi}{\partial R} = \left(ik - \frac{1}{R}\right)\Phi, \tag{32.20}$$

the integrand in (32.17), and hence the integral itself, are indeed equal to zero, i.e. radiation condition (32.18) is satisfied.

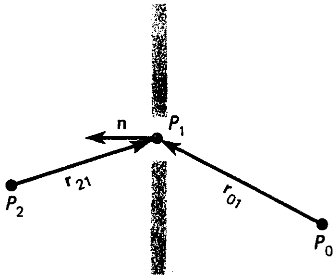


Fig. 157. To the derivation of the Fresnel-Kirchhoff diffraction relation.

KIRCHHOFF'S APPROXIMATION. In order to use formula (32.14) as a formula for computing $\Phi(P_0)$ from known values of Φ and $\partial\Phi/\partial n$ at points on a plane screen rather than an integral equation for determining Φ, Kirchhoff proposed the following rules for determining the values of $\Phi(P_0)$ in the plane of the screen.

1. At the apertures Φ and $\partial\Phi/\partial n$ have the same values that they would have in the absence of the opaque screen.
2. At the opaque parts of the screen $\Phi = 0$ and $\partial\Phi/\partial n = 0$.

The choice of the boundary conditions in accordance with these rules leads to the solution of the diffraction problems in Kirchhoff's approximation.

Kirchhoff's boundary conditions are never satisfied exactly: (a) at the aperture edges, certain boundary conditions must be satisfied (in principle, these conditions can be found from the electromagnetic theory of light); (b) there cannot be a sharp shadow behind the screen, i. e. Φ cannot abruptly turn to zero.

However, if the linear size of the apertures is much larger than the wavelength, the edge effects can be neglected, and Kirchhoff's boundary conditions provide a good approximation to the real situation.

OPTICAL APPROXIMATION. Since $\lambda \simeq 1$ nm in the visible region, the following condition is satisfied in practically all cases of any importance:

$$1/r_0 \ll k. \tag{32.21}$$

This condition is called the optical approximation condition. When this condition is satisfied, we can put

$$\frac{\partial}{\partial n}\left(\frac{\mathrm{e}^{ikr_{01}}}{r_{01}}\right) = \cos(\mathrm{n}, \mathrm{r}_{01})\left(ik - \frac{1}{r_{01}}\right)\frac{\mathrm{e}^{ikr_{01}}}{r_{01}}$$
$$\approx ik\cos(\mathrm{n}, \mathrm{r}_{01})\,\mathrm{e}^{ikr_{01}}/r_{01}. \tag{32.22}$$

Formula (32.14) can then be presented in the form

$$\Phi(P_0) = \frac{1}{4\pi}\int_S \frac{\mathrm{e}^{ikr_{01}}}{r_{01}}$$
$$\times\left[\frac{\partial\Phi}{\partial n} - ik\Phi\cos(\mathrm{n}, \mathrm{r}_{01})\right]\mathrm{d}S. \tag{32.23}$$

FRESNEL-KIRCHHOFF DIFFRACTION RELATION. Suppose that a spherical wave emanating from point P_2 (Fig. 157) is incident on an aperture:

$$\Phi(P_1) = A\mathrm{e}^{ikr_{12}}/r_{12}. \tag{32.24}$$

Considering that

$$\partial r_{12}/\partial n = \cos(\mathrm{n}, \mathrm{r}_{21}) = -\cos(\mathrm{n}, \mathrm{r}_{12}), \tag{32.25}$$

!

The application of Green's formula does not lead to a solution of the wave-field problem but simply provides an integral equation for determining the wave field. Kirchhoff's approximation is applicable in view of the fact that the wavelength of light is small as compared to the linear dimensions of screens and apertures in them.

?

What are the conditions under which Green's formula can be applied?
What is the physical meaning of the radiation condition?
What are the conditions for the applicability of Fresnel's approximation?

we obtain in the optical approximation

$$\partial\Phi/\partial n = -ik\cos(\mathrm{n},\mathrm{r}_{12})\,\mathrm{e}^{ikr_{12}}/r_{12}. \tag{32.26}$$

Consequently, formula (32.23) assumes the form

$$\Phi(P_0) = -A\frac{ik}{4\pi}\int_{S_0}\frac{\mathrm{e}^{ik(r_{01}+r_{12})}}{r_{01}r_{12}}[\cos(\mathrm{n},\mathrm{r}_{12}) + \cos(\mathrm{n},\mathrm{r}_{01})]\,\mathrm{d}S \tag{32.27}$$

(S_0 is the area of the aperture), since the integrand is equal to zero on the opaque parts of the screen. Equality (32.27) is called the Fresnel-Kirchhoff diffraction relation.

HELMHOLTZ' RECIPROCITY THEOREM. The sum of the cosines in the integrand of (32.27) does not vary if the point source and the point of observation are interchanged. This means that a point source situated at P_2 produces the same effect at point P_0 as the effect produced at point P_2 by a point source of the same intensity situated at P_0. This statement is called Helmholtz' reciprocity theorem.

SECONDARY SOURCES. Formula (32.27) can be written in the form

$$\Phi(P_0) = \int_{S_0}\Phi'(P_1)\frac{\mathrm{e}^{ikr_{01}}}{r_{01}}\,\mathrm{d}S, \tag{32.28}$$

where

$$\Phi'(P_1) = -A\frac{ik}{4\pi}\frac{\mathrm{e}^{ikr_{12}}}{r_{21}}[\cos(\mathrm{n},\mathrm{r}_{12}) + \cos(\mathrm{n},\mathrm{r}_{01})]. \tag{32.29}$$

According to the Huygens-Fresnel principle, the field at point P_0 is produced by secondary sources on the aperture S_0, the intensity of these sources being characterized by formula (32.29). It can be seen that these sources differ from the wave $A\exp(ikr_{12})/r_{12}$ in the following: (1) the amplitude of the secondary wave differs from the amplitude of the incident wave by a factor $k/(4\pi)$; (2) the dependence of the secondary wave amplitude on the direction of propagation is given by the factor $\cos(\mathrm{n},\mathrm{r}_{12}) + \cos(\mathrm{n},\mathrm{r}_{01})$, which is different from the factor proposed by Fresnel; (3) the phase of the secondary source differs from that of the incident wave by $\pi/2$ due to the presence of the multiplier $-i$. This solves the problem of the wave phase mentioned in connection with the Fresnel zone method at the end of Sec. 31.

Thus, within the framework of the electromagnetic theory of light, the Kirchhoff approximation allows us to overcome the difficulties associated with the Fresnel zone method and answer the questions that were beyond the scope of Fresnel's theory (the amplitude of the secondary sources, the nature of the dependence of the amplitude of secondary source radiation on the direction of propagation and the phase).

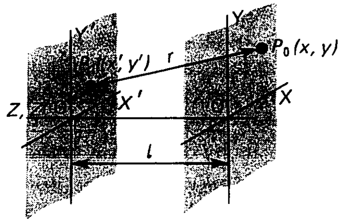


Fig. 158. Arrangement of coordinate systems in the planes of the source and diffraction pattern.

FRESNEL'S APPROXIMATION. Under typical conditions, the diffraction pattern is observed in a plane parallel to the screen with apertures in it. The plane in which the diffraction pattern is observed is called the diffraction pattern plane, while the other plane is called the source plane. In each plane, we introduce rectangular Cartesian coordinate systems whose *X*- and *Y*-axes are parallel, and the *Z*-axes coincide. The positive direction of the *Z*-axis is related to the directions of the *X*- and *Y*-axes in the usual manner (Fig. 158). The unit normal vector [see (32.27)] points in the direction from which the radiation is propagating. The function $e^{ikr_{01}}/r_{01}$ describes a wave moving towards point P_0. We denote by (x, y) the coordinates of point P_0 in the diffraction pattern plane, and by (x', y') the coordinates of the point of integration P_1 in the source plane; $dS' = dx'dy'$ is an area element on the source surface, and

$$\Psi(x', y') = Ae^{ikr_{12}}/r_{12} \tag{32.30}$$

is the amplitude of the sources. Then [see (32.27)]

$$\Phi(x, y) = \frac{-ik}{4\pi} \int_{S'} \Psi(x', y') \frac{e^{ikr}}{r} [\cos(\mathrm{n}, \mathrm{r}_{12}) + \cos(\mathrm{n}, \mathrm{r}_{01})] \, dS, \tag{32.31}$$

where

$$r = \sqrt{l^2 + (x - x')^2 + (y - y')^2}, \tag{32.32}$$

and *l* is the distance between the planes. The term with cosines in the integrand is a slowly varying function in comparison with the rapidly oscillating factor exp (*ikr*). It has practically no effect on the interference pattern, does not change its visibility and causes only an insignificant variation of the average brightness. Moreover, in most cases of practical importance, the angles $(\mathrm{n}, \mathrm{r}_{12})$ and $(\mathrm{n}, \mathrm{r}_{01})$ vary within small intervals near the zero value. Under this assumption, called the small angle approximation, we can put the cosines equal to unity without significantly altering the results:

$$\cos(\mathrm{n}, \mathrm{r}_{12}) = 1, \quad \cos(\mathrm{n}, \mathrm{r}_{01}) = 1. \tag{32.33}$$

Formula (32.31) can then be presented in a form convenient from the point of view of practical applications:

$$\Phi(x, y) = \frac{k}{2\pi i} \int_{S'} \Psi(x', y') (e^{ikr}/r) \, dx'dy'. \tag{32.34}$$

It is possible to take $\cos(\mathrm{n}, \mathrm{r}_{12}) + \cos(\mathrm{n}, \mathrm{r}_{01})$ out of the integral by using Eqs. (32.33) due to the fact that this function varies much more slowly than the rapidly oscillating exponential factor $\exp(ikr)$. Strictly speaking, while taking this function out of the integral, we must replace it by a certain value defined by the mean-value theorem. However, the exact value of this quantity is not important since, being a multiplier, it does not affect the visibility of the interference pattern and serves only as a scale factor. In most of the cases of practical interest, there is no need to exactly calculate the intensities of the diffraction patterns. Hence there is no point in unnecessarily complicating the formulas, and equalities (32.33) can be used without any loss of generality.

Further simplification of the formula is also made with the help of the small angle approximation. Mathematically, this approximation can be formulated in the form of the inequalities

$$|x - x'|/l \ll 1, \quad |y - y'|/l \ll 1. \tag{32.35}$$

If condition (32.35) is satisfied, expression (32.32) can be expanded into a series in (32.35). Confining to the second-order terms, we can write

$$\begin{aligned} r &= l\,[1 + (x - x')^2/l^2 + (y - y')^2/l^2]^{12} \\ &= l + \frac{(x - x')^2 + (y - y')^2}{2l}. \end{aligned} \tag{32.36}$$

Substituting expansion (32.36) into (32.34), we obtain

$$\begin{aligned} \Phi(x, y) &= \frac{k}{2\pi i}\frac{e^{ikl}}{l}\int_{S'} \Psi(x', y') \\ \exp &\times \left\{\frac{ik\,[(x - x')^2 + (y - y')^2]}{2l}\right\} dx'dy', \end{aligned} \tag{32.37}$$

where the slowly varying quantity $r \approx l$ in the denominator is taken out of the integral since it does not affect the visibility of the interference pattern and has only a slight influence on its overall brightness.

In most cases of practical importance, the diffraction can be described quite accurately with the help of approximation (32.37), called Fresnel's approximation. We shall term the diffraction considered in this approximation the Fresnel diffraction.

Under certain conditions, we can further simplify formula (32.37) and go over to the Fraunhofer approximation.

SEC. 33. FRAUNHOFER DIFFRACTION

Diffraction in the far region is considered and the criterion for the far region is discussed. A diffraction grating is studied and diffraction at periodic structures and inhomogeneous media is described.

FRAUNHOFER DIFFRACTION REGION. Expressing the exponent in (32.37) in the form

$$\frac{ik[(x-x')^2+(y-y')^2]}{2l}=\frac{ik(x^2+y^2)}{2l}+\frac{ik(x'^2+y'^2)}{2l}-\frac{ik(xx'-yy')}{l}, \tag{33.1}$$

we obtain

$$\Phi(x,y)=\frac{k}{2\pi i}\frac{e^{ikl}}{l}exp\left[\frac{ik(x^2+y^2)}{2l}\right]\times\int_{S'}\Psi(x',y')\exp\left[\frac{ik(x'^2+y'^2)}{2l}\right]\times\exp\left[-\frac{ik(xx'+yy')}{l}\right]\mathrm{d}x'\mathrm{d}y'. \tag{33.2}$$

The intensity distribution in the diffraction pattern is determined by the square of the modulus of $\Phi(x, y)$. Hence the exponents in front of the integral in (33.2) do not affect the intensity distribution in the diffraction pattern as their moduli are equal to unity. The integration in (33.2) is carried out over the entire area S' of the plane, considering that $\Psi(x', y') = 0$ at the points on the opaque parts of the screen. In other words, integration with respect to x' and y' is carried out from $-\infty$ to ∞. Hence, to within the multipliers, one of which is a phase multiplier depending on (x, y), $\Phi(x, y)$ is a Fourier-function transform:

$$\Psi(x',y')\exp[ik(x'^2+y'^2)/(2l)]\ldots \tag{33.3}$$

In the mathematical sense, the study of diffraction is reduced to the application of the Fourier transformation theory.

The presence of the exponential factor in (33.3) somewhat complicates the task of the Fourier transformation and hence makes an analysis of the diffraction pattern difficult. The simplest case of diffraction is realized by getting rid of this factor. If the aperture is quite small and the distance to the diffraction pattern plane is quite large ($l \to \infty$), we get

$$\exp[ik(x'^2+y'^2)/(2l)]\to 1. \tag{33.4}$$

Diffraction under this condition is called the Fraunhofer diffraction.

However, the exponential factor in (33.3) can be neglected not only for infinite values of l. It is sufficient that this term should not oscillate, i. e. the exponent must not exceed $\pi/2$. For example, $\cos(\pi/4) = \sin(\pi/4) = 0.7$, and for the criterion determining whether the effect of the exponential factor in (33.3) on diffraction can be neglected, we can take the condition

$$\frac{k(x'^2 + y'^2)}{2l} = \frac{k\rho'^2}{2l} < \frac{\pi}{4}, \tag{33.5}$$

where

$$\rho'^2 = x'^2 + y'^2 \tag{33.6}$$

is the square of the maximum distance from the centre to the edge of the aperture at which diffraction takes place. From (33.6) we conclude that the Fraunhofer diffraction is observed at distances

$$l > l_{\min} = 2k\rho'^2/\pi = 4\rho'^2/\lambda. \tag{33.7}$$

In other words, the region of Fraunhofer diffraction extends from a certain minimum distance [see (33.7)] to infinity. For example, when $\lambda \simeq 0.5\,\mu\text{m}$, $\rho' = 1\text{ cm} = 10^{-2}\text{ m}$, we obtain $l_{\min} = 800\text{ m}$, i. e. the Fraunhofer diffraction can be observed from quite large distances. However, even for $\rho' = 1\text{ mm} = 10^{-3}\text{ m}$, we obtain $l_{\min} = 8\text{ m}$, which is closer to the laboratory conditions. For $\rho' = 0.1$ mm we have $l_{\min} = 8$ cm, and the Fraunhofer diffraction can be observed in a very small region.

The factor in front of the integral in (33.2) does not affect the intensity distribution in the diffraction pattern. For convenience of writing the formulas for the Fraunhofer diffraction, it is expedient to reduce this factor to unity. Formula (33.2) then assumes the form

$$\boxed{\Phi(x, y) = \int_{S'} \Psi(x', y') \exp\left[-\frac{ik(xx' + yy')}{l}\right] \mathrm{d}x'\mathrm{d}y'.} \tag{33.8}$$

It should be remembered that in this form of notation, formula (33.8) can be used for computing relative, and not absolute, intensities in the diffraction pattern. Moreover, Φ and Ψ in the formula have different dimensions. For example, if Φ is the amplitude, then Ψ has the dimensions of amplitude divided by area, since the left- and right-hand sides of (33.8) should have the same dimensions.

It can be seen from (33.8) that the diffraction pattern depends on x/l and y/l. This means that the diffraction pattern is magnified without distortion in proportion to the distance.

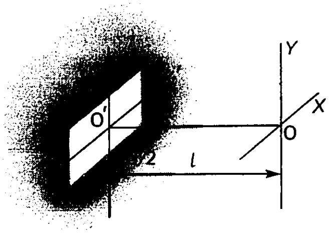


Fig. 159. To the calculation of diffraction at a rectangular aperture.

Considering that the Fraunhofer diffraction is realized in pure form at infinity ($l \to \infty$), we conclude that this type of diffraction takes place in parallel rays. Hence the term parallel-ray diffraction is sometimes used for the Fraunhofer diffraction.

In a region satisfying condition (33.7), the Fraunhofer diffraction can be observed directly on the screen without any additional devices. However, it is more convenient to observe the diffraction pattern in the focal plane of a converging lens placed in the path of diffracted rays.

DIFFRACTION AT A RECTANGULAR APERTURE. We assume that the origin of coordinates is located at the centre of a rectangular aperture with sides a and b (Fig. 159). A plane wave directed along the Z-axis is incident on the aperture, where its phase and amplitude are constant. We denote by A_0 the complex wave amplitude at the aperture. Considering that Ψ' in (33.8) has the dimensions of amplitude divided by area, we put

$$\Psi' = A_0/(ab). \tag{33.9}$$

Then, in accordance with formula (33.8), we get

$$A(x, y) = (A_0/(ab)) \int_{-a/2}^{a/2} \exp(-ikxx'/l)\,dx' \times \int_{-b/2}^{b/2} \exp(-ikyy'/l)\,dy', \tag{33.10}$$

where $A(x, y) = \Phi(x, y)$ is the amplitude. Elementary integration in (33.10) leads to the formula

$$A = A_0 (\sin\alpha/\alpha)(\sin\beta/\beta), \tag{33.11}$$

where

$$\alpha = kax/(2l) \text{ and } \beta = kby/(2l). \tag{33.12}$$

The observed intensity is proportional to $|A|^2$ and, but for a constant multiplier, can be presented in the form

$$I(x, y) = |A_0|^2 \frac{\sin^2\alpha}{\alpha^2} \frac{\sin^2\beta}{\beta^2}. \tag{33.13}$$

The zero-intensity fringes are defined by the condition $\sin\alpha = 0$, $\sin\beta = 0$, and are straight lines parallel to the X- and Y-axes forming a system of rectangles. Inside the rectangles, the intensity is nonzero (Fig. 160). The dependence of the factor $\sin^2\alpha/\alpha^2$ on α is represented in Fig. 161. In addition to the central peak, other peaks are observed at the values $\alpha = 3\pi/2$, $5\pi/2$, etc. Hence the intensity in successive peaks decreases in the proportion $1:[2/(3\pi)]^2:[2/(5\pi)]^2:\ldots = 1:0.04:0.016:\ldots$. In other words, the intensity drops very rapidly. Hence most of the energy of light transmitted through an aperture is concentrated in the central bright rectangular spot.

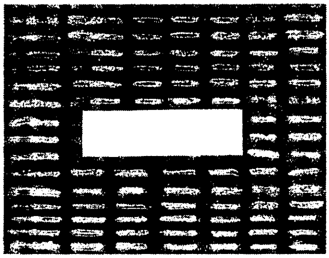


Fig. 160. Diffraction at a rectangular aperture.

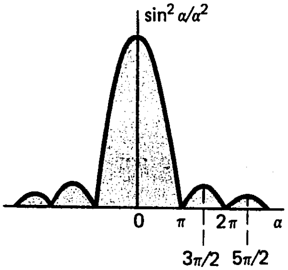


Fig. 161. Curve describing the function $\sin^2\alpha/\alpha^2$.

It follows from (33.12) that α and β depend only on the ratios x/l and y/l. This means that the size of the diffraction pattern increases in proportion to the distance from the aperture, i. e. from the centre of the aperture the diffraction pattern is seen at a constant angle (Fig. 162). In accordance with Fraunhofer's approximation, this angle is assumed to be small. Hence α and β can be transformed as follows:

$$\alpha = kax/(2l) = (ka/2)\tan\varphi_x = (ka/2)\sin\varphi_x,$$

$$\beta = (ka/2)\sin\varphi_y, \qquad (33.14)$$

where $\tan\varphi_x = \sin\varphi_x \approx x/l$, $\sin\varphi_y \approx y/l$ since the angles φ_x and φ_y are small. Notation (33.14) shows that the diffraction pattern really depends only on the angle between the ray and the Z-axis, i. e. $A(x, y) = A(\varphi_x, \varphi_y)$. For example, Fig. 162 shows the rays in the plane $X = 0$. In this case, $\varphi_x = 0$, $\varphi_y = \varphi$ and $A = A(\varphi)$. Consequently, the diffracted rays emanating from different points of the aperture and forming a diffraction pattern at infinity at an angle φ are parallel and inclined at an angle φ to the axis. In order to obtain an interference pattern at a finite distance, we must make the parallel rays converge at a point. For this purpose, we must place a converging lens in the path of these rays. The diffraction pattern will then be observed in the focal plane of this lens.

DIFFRACTION AT A SLIT. If we increase the length of a rectangular aperture, the period of repetition of diffraction fringes perpendicular to the longer side decreases. If the length of the aperture is quite large, this period becomes so small that the eye is notable to distinguish between the fringes. The diffraction pattern then contains fringes parallel to the longer side of aperture, and is called the diffraction pattern formed by a slit (Figs. 163, 164). In this case, formula (33.11) can be written in the form

$$A_\varphi = A_0 \sin\beta/\beta, \qquad (33.15)$$

where $\beta = kby/(2l) = kb\sin\varphi/2$. Formula (33.15) can be derived with the help of the Fresnel zone method. We take the phase of the wave at the centre of the slit as the reference phase for the diffracted waves. The wave from point y' (Fig. 165) propagates in a direction determined by the angle φ and acquires a phase difference $\delta = k\Delta = ky'\sin\varphi$ in comparison with the wave at point O. Obviously, the amplitude of the wave per slit element of width dy' is equal to $A_0 dy'/b$. Hence

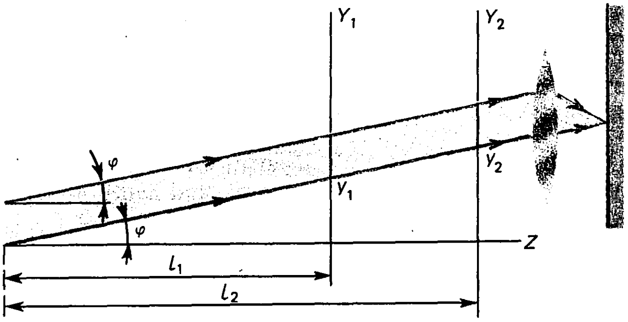


Fig. 162. Fraunhofer diffraction: diffraction in parallel rays.

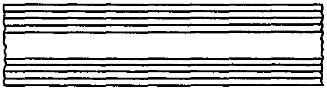


Fig. 163. Diffraction at a slit (the fringes are parallel to the edges of the slit).

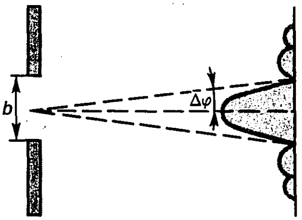


Fig. 164. Intensity distribution in the case of diffraction at a slit.

the complex amplitude obtained as a result of superposition of these waves at infinity is given by

$$A_\varphi = (A_0/b) \int_{-b/2}^{b/2} \mathrm{e}^{-iky' \sin\varphi}\, \mathrm{d}y' = A_0 \frac{\sin\beta}{\beta},$$
$$\beta = kb \sin\varphi/2 \tag{33.16}$$

[see (33.15)]. A converging lens must be used in order to observe diffraction pattern at a finite distance.

The angle $\Delta\varphi$ at which the first dark fringe is seen from the centre of the slit is determined from the condition

$$\beta = kb \sin\Delta\varphi/2 = \pi. \tag{33.17}$$

Assuming $\Delta\varphi$ to be small and putting $\sin\Delta\varphi \approx \Delta\varphi$, we find that the angle at which the central bright fringe is seen from the centre of the slit is

$$\theta_0 = 2\Delta\varphi = 2\lambda/b. \tag{33.18}$$

As the slit width is reduced, the angular size of the bright fringe increases, and for $b \to \lambda$ the intensity decreases uniformly from the centre to the periphery without any deviations.

DIFFRACTION AT A CIRCULAR APERTURE. Let us denote by a the radius of a circular aperture in the $X'Y'$ plane with its centre coinciding with the origin of coordinates (Fig. 166). It is convenient to use polar coordinates to make calculations in the source plane as well as in the diffraction plane:

$$x' = r' \cos\theta', \; x = r\cos\theta, \; y' = r'\sin\theta', \; y = r\sin\theta. \tag{33.19}$$

Considering that

$$\mathrm{d}x'\mathrm{d}y' = r'\mathrm{d}r'\mathrm{d}\theta', \; \sin\theta\sin\theta' + \cos\theta\cos\theta' = \cos(\theta' - \theta), \tag{33.20}$$

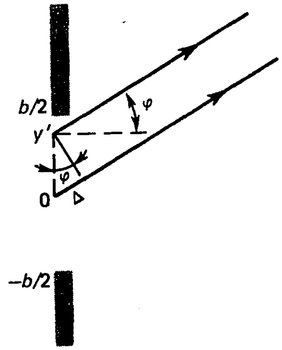


Fig. 165. To the calculation of diffraction at a slit by the Fresnel zone method.

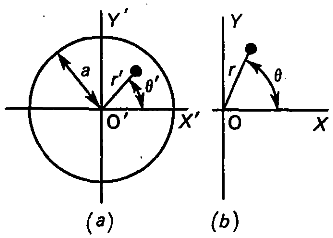


Fig. 166. Polar coordinate systems in the planes of the source and the diffraction pattern.

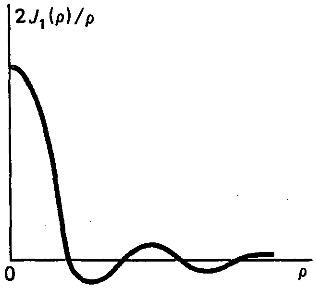


Fig. 167. Curve of the function $J_1(\rho)/\rho$, determining the diffraction pattern at a circular aperture.

we can write formula (33.8) as

$$A(r, \theta) = (A_0/(\pi a^2)) \int\limits_0^a r' \mathrm{d}r' \int\limits_\theta^{2\pi+\theta} \exp\left[ikr'(r/l) \cos(\theta - \theta')\right] \mathrm{d}\theta', \tag{33.21}$$

where it is assumed that the wave front of the plane wave incident on a circular aperture is parallel to the plane of the aperture and hence $\Psi = A_0/(\pi a^2)$. The inner integral is equal to

$$\int\limits_\theta^{2\pi+\theta} \exp\left[i\eta \cos(\theta - \theta')\right] \mathrm{d}\theta' = \int\limits_0^{2\pi} \exp\left[i\eta \cos \theta'\right] \mathrm{d}\theta' = 2\pi J_0(\eta),$$

where $J_0(\eta)$ is the zeroth-order Bessel function, and $\eta = krr'/l$. Then [see (33.21)]

$$A = (2A_0/a^2) \int\limits_0^a r' \mathrm{d}r' J_0(krr'/l). \tag{33.22}$$

From the theory of Bessel functions, we get the relation

$$\int x J_0(x) \,\mathrm{d}x = x J_1(x), \tag{33.23}$$

where $J_1(x)$ is the first-order Bessel function. Taking (33.23) into account, we obtain

$$\int\limits_0^a r' J_0(krr'/l) \,\mathrm{d}r' = [la/(kr)]\, J_1(kra/l). \tag{33.24}$$

This gives [see (33.22)]

$$A(r) = A_0 \frac{2J_1(kar/l)}{kar/l}. \tag{33.25}$$

Bessel functions have been studied quite thoroughly. The dependence of the function $2J_1(\rho)/\rho$ on ρ is shown in Fig. 167. There is a bright circular spot at the centre of the diffraction pattern. This spot is surrounded by alternating bright and dark rings. The intensity maxima in bright rings rapidly attenuate (Fig. 168). The radius r_1 of the first dark ring is determined from the condition

$$J_1(kar_1/l) = 0. \tag{33.26}$$

Considering that the first root of the equation $J_1(x) = 0$ is $x_1 = 3.832$, we can determine the angle $\Delta\varphi_1$ at which this ring is seen from the centre of the circular aperture:

$$\Delta\varphi_1 = (r_1/l) = 3.832/(ka) = 0.61\lambda/a. \tag{33.27}$$

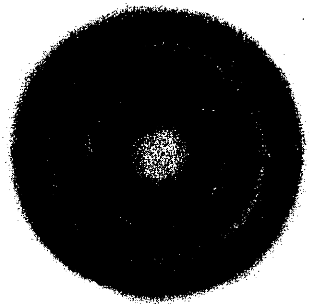


Fig. 168. Diffraction at a circular aperture.

Hence the angular size of the bright spot seen from the centre of the circular aperture is

$$\theta_0 = 2\Delta\varphi_1 = 1.22\lambda/a. \tag{33.28}$$

This formula plays a significant role in determining the resolving power of optical instruments (see Sec. 36). The second ring is seen at an angle

$$\Delta\varphi_2 = 1.116\lambda/a. \tag{33.29}$$

Outside the central bright spot, however, the intensity decreases very slowly. About 84% of the total energy passing through the aperture is concentrated in the central bright spot. In most cases, the energy in the region outside the central spot can be neglected.

DIFFRACTION GRATING. A diffraction grating is an assembly of periodically repeated slits (Fig. 169). We denote by a, b and $d = a + b$ the width of the transparency, the opaque part of the slit and the grating period respectively. The diffraction pattern can be obtained with the help of formula (33.8) in the same way as in the case of a single slit. However, the composition of complex amplitudes is a convenient and graphic method of analysis in this case.

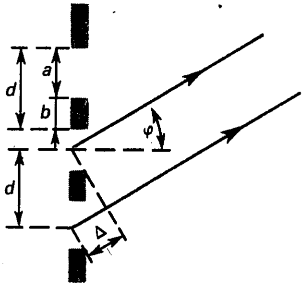


Fig. 169. Diffraction grating.

The amplitude of a wave diffracted at each slit in the direction characterized by angle φ [see (33.15)] can be presented in the form

$$A_\varphi^{(0)} = A_0 \sin\beta/\beta, \quad \beta = kb\sin\varphi/2, \tag{33.30}$$

where the superscript (0) on A_φ indicates that this is the wave amplitude at a single slit. The waves diffracted at the slits interfere and form a diffraction pattern. Thus the diffraction pattern formed by a grating is the result of diffraction of waves from various slits and the interference of waves diffracted by these slits.

Let us consider the interference of waves from the slits. The path difference for waves from two adjacent slits (Fig. 169) and the phase difference between them are

$$\Delta = d\sin\varphi, \tag{33.31}$$

$$\delta = k\Delta = kd\sin\varphi. \tag{33.32}$$

Taking the phase of the wave diffracted at an angle φ from the first slit as the reference point, i.e. taking its amplitude equal to (33.30), we obtain the expressions $A_\varphi^0 e^{-i\delta}$, $A_\varphi^0 e^{-2i\delta}$, ... for the amplitudes of the waves diffracted at the second, third, ... slits. The superposition of waves from all the N slits results in the formation of a wave with the amplitude

$$A_\varphi = A_\varphi^{(0)} + A_\varphi^{(0)} e^{-i\delta} + A_\varphi^{(0)} e^{-i2\delta} + \dots + A_\varphi^{(0)} e^{-(N-1)i\delta} = A_\varphi^{(0)}(1 - e^{-iN\delta})/(1 - e^{-i\delta}), \tag{33.33}$$

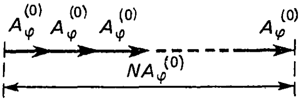


Fig. 170. Composition of complex amplitudes leading to the formation of principal maxima.

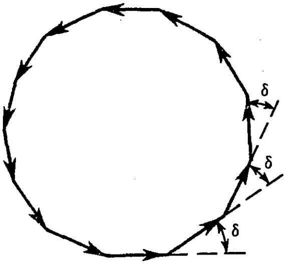


Fig. 171. Composition of complex amplitudes leading to the formation of minima.

where we have used the following formula for geometric progression:

$$1 + x + x^2 + \ldots + x^n = (1 - x^{n+1})/(1 - x).$$

Considering that

$$\frac{1 - e^{-iN\delta}}{1 - e^{-i\delta}} = \frac{e^{-iN\delta/2}}{e^{-i\delta/2}} \frac{e^{iN\delta/2} - e^{-iN\delta/2}}{e^{i\delta/2} - e^{-i\delta/2}} = \frac{e^{-iN\delta/2}}{e^{-i\delta/2}} \frac{\sin(N\delta/2)}{\sin(\delta/2)}, \tag{33.34}$$

and using (33.30), we finally obtain

$$|A_\varphi| = |A_0|\,|\sin\beta/\beta|\,|\sin N\alpha/\sin\alpha|, \quad \beta = kb\sin\varphi/2,$$
$$\alpha = \delta/2 = kd\sin\varphi/2. \tag{33.35}$$

This formula completely describes the diffraction pattern formed by a slit, since the intensity of the pattern is given by $|A_\varphi|^2$. It can be seen from (33.31) that for

$$d\sin\varphi = m\lambda \quad (m = 0, \pm 1, \pm 2, \ldots) \tag{33.36}$$

the waves from adjacent slits augment each other, i.e. waves from all slits are mutually augmented. This means that relation (33.36) determines the directions in which the principal maxima are formed. Under this condition, $\alpha = kd\sin\varphi/2 = \pi m$, and hence

$$|\sin N\alpha/\sin\alpha| = N. \tag{33.37}$$

The composition of complex amplitudes from individual slits, leading to the formation of principal maxima, is shown graphically in Fig. 170. It can be seen that $N|A_\varphi^{(0)}|$ is indeed the maximum amplitude which can be obtained from the amplitudes of the waves diffracted at N slits in the direction determined by the angle φ. That is why these maxima are called the principal maxima. However, $|A_\varphi|$ of the amplitudes of principal maxima is not identical. It can be seen from (33.35) that it is modulated by a factor $|\sin\beta/\beta|$, i.e. the amplitude of the principal maxima is modulated by diffraction at individual slits. The maximum value of $|\sin\beta/\beta|$ is equal to unity and is attained under the condition $\beta = 0$, which corresponds to the central maximum ($\varphi = 0$). The amplitudes of all the remaining principal maxima are lower than this value. If the principal maximum corresponds to the direction for which $\sin\beta = 0$, then it will not be observed. Such a situation arises when b and d are commensurate.

The minima are formed when the composition of complex amplitudes results in zero amplitude (Fig. 171). For different values of δ (i.e. for different φ), the broken line can be closed

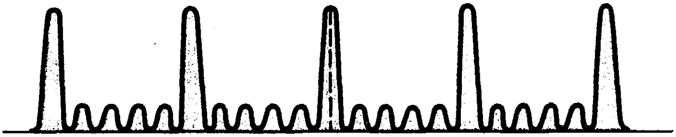


Fig. 172. General nature of the diffraction pattern at six slits (not to scale).

once, twice, and so on, i. e. the phase difference between waves from the extreme slits in the grating is equal to 2π, 4π, ... Hence the condition for obtaining minimum amplitudes in the diffraction pattern is written in the form

$$N\delta = 2\pi n \quad (n = 0, 1, 2, \ldots) \tag{33.38}$$

or, since $\delta = kd \sin \varphi = 2\pi d \sin \varphi/\lambda$,

$$d \sin \varphi = (n/N)\lambda \quad (n \neq 0, N, 2N, \ldots), \tag{33.39}$$

since for $n = 0, N, 2N, \ldots$ the ratio n/N becomes equal to an integer ($m = n/N$) and (33.39) is transformed into the condition for the principal maxima. For the minima, condition (33.38) can be obtained from (33.35):

$$\sin (N\delta/2) = 0. \tag{33.40}$$

Hence $N\delta/2 = n\pi$ ($n = 0, 1, 2, \ldots$), i. e. coincides with (33.38). If $n/N = m$, then $\sin N\alpha/\sin \alpha$ remains finite and leads to the formation of the principal maximum.

It can be concluded from (33.39) that there are $N - 1$ minima between two principal maxima. It is clear that between the minima, there exist maxima which are called secondary maxima. Hence there are $N - 2$ secondary maxima between two principal maxima. If the number N of slits is known, we can determine the general nature of the diffraction pattern, while formula (33.35) can be used to determine the relative moduli of amplitudes and intensities of different parts of the diffraction pattern.

The general nature of the distribution of the absolute value of the amplitude for six slits ($N = 6$) is shown in Fig. 172 (not to scale).

DIFFRACTION OF WHITE LIGHT AT A GRATING. Waves of different wavelengths form their own diffraction patterns. It can be seen from the condition $d \sin \varphi = m\lambda$ that for a fixed m, the angle φ increases with λ. In the zero-order interference ($m = 0$), the central maximum $\varphi = 0$ is the same for all waves. Hence a white fringe is formed at the centre. This is followed by the first-order interference ($m = 1$). The interference fringes of the first and second orders do not overlap. Overlapping of interference fringes of different orders takes place when $m\lambda = m'\lambda'$, where λ and λ' are the wavelengths and m and m' are the orders of their interference. For the visible spectrum, we have

$$1 \times 0.76 \times 10^{-6} = 2 \times 0.38 \times 10^{-6}, \tag{33.41}$$

i.e. the long-wave boundary of the first spectrum coincides with the very beginning of the short-wave boundary of the second-order spectrum. Hence the first-order visible part of the spectrum does not overlap with the second-order spectrum. The second- and third-order spectra overlap. It can be seen from (33.41) that this circumstance is due to the visible part of the spectrum having only one octave.

DISPERSION REGION. This region for a diffraction grating is obtained in the same way as in the case of the Fabry-Perot interferometer (see Sec. 28) and is determined from formula (28.34). Spectra of lower orders ($m = 1, 2, 3, \ldots$) are usually observed in a diffraction grating, and hence the diffraction region is quite large ($\Delta\lambda = \lambda$, $\Delta\lambda = \lambda/2$, $\Delta\lambda = \lambda/3$). In particular, the first-order dispersion region for a diffraction grating coincides with the entire visible region of the spectrum.

RESOLVING POWER. For a diffraction grating, the resolving power is determined in the same way as for the Fabry-Perot interferometer (see Sec. 28). However, the resolution condition for fringes is now taken in the form of the Rayleigh condition: fringes are assumed to be resolved if the intensity maximum of one fringe falls on the intensity minimum of the second fringe.

It can be seen from (33.39) that the principal maximum appears for $n = mN$, and the adjoining minimum is observed at $n = mN + 1$.

Consequently, we can write the following relations from (33.39) for the corresponding angles:

$$d \sin \varphi_{\max} = mN\lambda/N = m\lambda, \tag{33.42}$$
$$d \sin \varphi_{\min} = (mN + 1)\lambda/N = (m + 1/N)\lambda.$$

The condition for the m-th order principal maximum has the following form for the wavelength $\lambda_1 = \lambda + \Delta\lambda$:

$$d \sin \varphi_{\max}^{(1)} = m\lambda_1. \tag{33.43}$$

In accordance with Rayleigh's condition, two fringes are considered to be resolved if $\varphi_{\min} = \varphi_{\max}^{(1)}$. Taking (33.42) and (33.43) into consideration, we can write this condition in the form

$$(m + 1/N)\lambda = m(\lambda + \Delta\lambda). \tag{33.44}$$

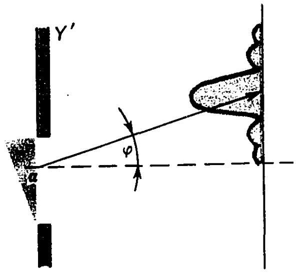


Fig. 173. Diffraction at a slit for a continuously varying wave phase.

This leads to the following expression for the resolving power of a grating [see (28.22)]:

$$\Upsilon = \lambda/\Delta\lambda = Nm. \tag{33.45}$$

Since there are tens of thousands of rulings ($N \sim 10^4$) in a diffraction grating, its resolving power is very high. In order to increase the resolving power of a grating, we must go over to large interference orders m. However, m is restricted by the angles of inclination $\varphi \leqslant \pi/2$, and $\sin\varphi$ in (33.36) cannot be more than unity. Hence $m_{\max} = d/\lambda$, and

$$\Upsilon_{\max} = Nd/\lambda, \tag{33.46}$$

is equal to the number of waves that can be accommodated over the entire width of the grating. Consequently, the maximum resolving power of a grating is independent of the number of rulings and of the distribution of the transparent and opaque parts of the grating over its width d. Gratings with large values of N have considerable advantages over those with small N in that they have a much better angular dispersion and a larger ratio of the variation $\delta\varphi$ of the angle upon a change in the wavelength to $\delta\lambda$, given by

$$D = \delta\varphi/\delta\lambda. \tag{33.47}$$

From (33.36), we obtain $d\cos\varphi\delta\varphi = m\delta\lambda$, and hence

$$D = m/(d\cos\varphi). \tag{33.48}$$

Hence a grating with a smaller d has a higher angular dispersion. In other words, a given value of angular dispersion is attained for this grating for a lower interference order m, which is very important since the intensity of maxima in lower orders is higher. Furthermore, the dispersion region for a small m is larger. Hence only diffraction gratings with a very large number N of periods and large total widths are of practical importance. The best types of gratings have a width of up to 15 cm and contain about 100 000 periods.

REFLECTION GRATINGS. If the transparencies on a reflection grating are replaced by highly reflecting regions and the opacities are replaced by regions that absorb light, the reflected light forms a diffraction pattern similar to the one observed in transmitted light. The gratings operating in reflected light are called reflection gratings. If the opacities of a grating are good reflectors of light and the transparencies are good transmitters, the grating can function in reflected as well as transmitted light. The period d of both gratings is the same while the quantities a and b have opposite meanings in the two cases and are interchanged in the formulas for intensity.

DIFFRACTION AT A SLIT FOR A CONTINUOUSLY VARYING WAVE PHASE. Let us assume that a prism with refractive index n is placed in a slit (Fig. 173). When a plane wave passes through a prism, the wave emerging from the slit acquires a phase shift which varies along the slit. As before, the origin of coordinates is assumed to be located at the centre of the slit. Assuming that the prism has a small apex angle α, we can take that at the point with coordinate y' the thickness of the prism is equal to $\alpha b/2 + \alpha y'$. The optical thickness is $(\alpha b/2 + \alpha y')n$, and the phase shift is

$$\delta = kn(\alpha b/2 + \alpha y') + k[\alpha b - (\alpha b/2 + \alpha y')], \tag{33.49}$$

where the second term takes into account the phase shift upon ray propagation over the path of length αb traversed by the ray in air. Hence Ψ in formula (33.8) has the following form for a single variable (the slit is assumed to be infinitely long):

$$\Psi(y') = (A_0/b)\,e^{i\delta}. \tag{33.50}$$

Hence the amplitude of the diffracted wave is [see (33.16)]

$$\begin{aligned} A_\varphi &= (A_0/b)\exp[ik(n+1)\alpha b/2] \\ &\times \int_{-b/2}^{b/2} \exp\{-i[k\sin\varphi - k(n-1)\alpha]y'\}\,dy' \\ &= A_0 e^{ik(n+1)\alpha b/2} \sin\beta'/\beta', \\ \beta' &= kb[\sin\varphi - (n-1)\alpha]/2. \end{aligned} \tag{33.51}$$

This gives

$$\begin{aligned} |A_\varphi| &= |A_0|\left|\frac{\sin\beta'}{\beta'}\right| \\ &= |A_0|\frac{\sin\{kb[\sin\varphi - (n-1)\alpha]/2\}}{kb[\sin\varphi - (n-1)\alpha]/2}. \end{aligned} \tag{33.52}$$

It can be seen that the central maximum of the diffraction pattern is located at a diffraction angle $\varphi = (n-1)\alpha$, and there are no significant variations in the diffraction apart from this (Fig. 173).

PHASE GRATINGS. If prisms are placed in all the slits of a diffraction grating as described above, the diffraction maximum will be displaced by the diffraction angle $\varphi = (n-1)\alpha$. The direction towards the principal maximum does not change since it is determined by the interference of waves from different slits, and the conditions under which this interference occurs remain unchanged (the path difference between the rays from adjacent slits is still $d\sin\varphi$). Hence the general form of the interference pattern remains the same and only the

!

Fresnel's diffraction takes place in the near zone, while Fraunhofer's diffraction is observed in the far zone. Fraunhofer's diffraction is described with the help of Fourier transforms.

intensity distribution changes: the maximum intensity cor**responds not to the principal maximum at the angle $\varphi = 0$, but to the principal maximum at the angle $\varphi = (n-1)\alpha$** or the one closest to this angle. This allows us to use higher-order interference, thus improving the resolving power and at the same time preventing any loss in intensity.

Gratings which cause a change in the phase of waves passing through them are called **phase gratings.** Diverse types of phase changes may take place along a slit. The appropriate value is chosen in accordance with the conditions of the problem which is solved with the help of the phase grating.

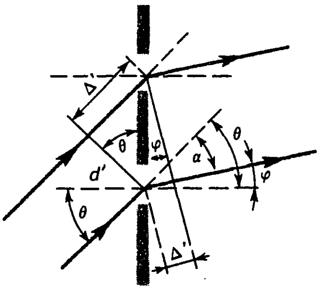


Fig. 174. Oblique incidence of rays on a grating.

AMPLITUDE-AND-PHASE GRATINGS. When light passes through a slit in such a grating, **the phase as well as amplitude of the waves change significantly. In other words, light is absorbed while passing through such gratings.**

OBLIQUE INCIDENCE OF RAYS ON A GRATING. When rays are obliquely incident on a grating (Fig. 174), the condition of principal maxima still remains the equality of the path difference between rays from adjacent slits to an integral multiple of wavelengths [see (33.36)]. It can be seen from Fig. 174 that the path difference between the rays is

$$\Delta = \Delta'' - \Delta' = d \sin \theta - d \sin \varphi .$$

Hence the condition for the principal maximum has the form

$$d(\sin \theta - \sin \varphi) - m\lambda . \tag{33.53}$$

We shall measure angles α from the direction of the incident ray. It can be seen from Fig. 174 that

$$\alpha = \theta - \varphi . \tag{33.54}$$

The diffraction angles are usually small, i.e. $\alpha \ll 1$, and hence

$$\begin{aligned} \sin \theta - \sin \varphi &= 2 \cos [(\theta + \varphi)/2] \sin [(\theta - \varphi)/2] \\ &= 2 \cos \theta \sin (\alpha/2) = \cos \theta \sin \alpha . \end{aligned} \tag{33.55}$$

Using the notation $d' = d \cos \theta$ (Fig. 174), we can write condition (33.53) in the form

$$d' \sin \alpha = m\lambda , \tag{33.56}$$

which is the same as (33.36). This means that **in the case of an oblique incidence of rays on a diffraction grating, the diffraction takes place in the same way as in the normal incidence, but the grating period is replaced by the projection of the grating period in a direction perpendicular to the incidence of the ray (Fig. 174). In other words, the diffraction as if occurs at a grating with a smaller period.**

This allows us to observe the diffraction from waves with a very small wavelength by using a large-period grating when

?

Under what conditions can Fraunhofer's diffraction be observed at short distances?
What are the reasons behind the large dispersion region of a diffraction grating?
Can you describe the diffraction at a grating by using the concepts of diffraction at continuously varying structures?

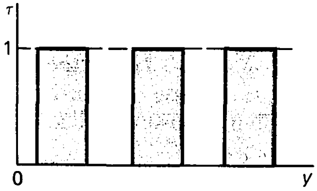


Fig. 175. Variation of the amplitude transmission coefficient for a diffraction grating.

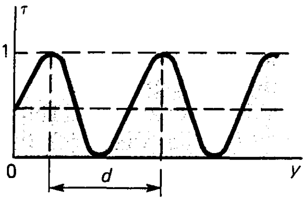


Fig. 176. A continuous structure with harmonically varying amplitude transmission coefficient.

practically no diffraction is observed when the rays are normally incident on the grating. It should be recalled that to observe the diffraction for normal incidence of the rays on the grating, we must ensure that the grating period is slightly larger than the wavelength. In an ordinary gramophone record, for example, the distance between the grooves in which the sound is recorded is about 40-50 μm. If the rays of white light are obliquely incident on the record, the effective separation between the grooves may become just a few micrometers, and the record will act like a reflection grating forming a first-order spectrum and coloured fringes in a higher-order interference.

DIFFRACTION AT CONTINUOUS PERIODIC AND APERIODIC STRUCTURES. A diffraction grating is a periodic structure with an amplitude transmission coefficient τ equal to 1 in the transparencies and 0 on the opacities (Fig. 175). If A_0 is the amplitude of the wave incident on a grating, the amplitude of the wave emerging from the grating is

$$A = A_0\tau = \begin{cases} A_0 & \text{for transparencies,} \\ 0 & \text{for opacities.} \end{cases}$$

Let us suppose that the grating is replaced by a layer of a material whose transmission coefficient varies according to the law (Fig. 176)

$$\tau(y') = (1/2)[1 + \cos(2\pi y'/d)]. \tag{33.57}$$

Moreover, it is assumed that the layer is infinite, i.e. $-\infty < y' < \infty$. If a wave with amplitude A_0 is incident on the layer, the amplitude of the wave leaving the layer is given by

$$A(y') = (A_0/2)\,[1 + \cos(2\pi y'/d)]. \tag{33.58}$$

In accordance with formula (33.8), we obtain the following distribution of amplitude A over the diffraction angles:

$$A_\varphi = \frac{B}{2\pi}\int_{-\infty}^{\infty} [1 + \cos(2\pi y'/d)]\, \mathrm{e}^{-iky'\sin\varphi}\, \mathrm{d}y', \tag{33.59}$$

where the constant B includes A_0 as well as a certain normalization factor which need not be determined more accurately since all calculations are carried out with an accuracy up to the relative amplitude variation. The factor $1/(2\pi)$ has been introduced for the sake of convenience. Considering that

$$\cos(2\pi y'/d) = (1/2)\,(\mathrm{e}^{i2\pi y'/d} + \mathrm{e}^{-i2\pi y'/d}), \tag{33.60}$$

we can present (33.59) in the form

$$A_\varphi = \frac{B}{2\pi}\int_{-\infty}^{\infty} \exp(-iky'\sin\varphi)\,dy'$$
$$+\frac{B}{4\pi}\int_{-\infty}^{\infty} \exp\left[i\left(\frac{2\pi}{d}-k\sin\varphi\right)y'\right]dy'$$
$$+\frac{B}{4\pi}\int_{-\infty}^{\infty} \exp\left[-i\left(\frac{2\pi}{d}+k\sin\varphi\right)y'\right]dy', \qquad (33.61)$$

whence

$$A_\varphi = B\delta(k\sin\varphi) + (B/2)\,\delta(2\pi/d - k\sin\varphi)$$
$$+ (B/2)\,\delta(2\pi/d + k\sin\varphi). \qquad (33.62)$$

Consequently, the diffraction pattern consists of three fringes. The fringe at the angle $\varphi = 0$ appears due to the constant component of the amplitude in (33.58) and corresponds to a simple rectilinear propagation of the wave through a homogeneous medium. Two fringes corresponding to the diffraction at angles $\varphi_{(+)}$ and $\varphi_{(-)}$ are determined from the conditions

$$2\pi/d - k\sin\varphi_{(+)} = 0 \text{ and } 2\pi/d + k\sin\varphi_{(-)} = 0. \qquad (33.63)$$

Considering that $k = 2\pi/\lambda$, we can present these conditions in the form

$$d\sin\varphi_{(+)} = \lambda \text{ and } d\sin\varphi_{(-)} = -\lambda. \qquad (33.64a)$$

Thus, only the first-order diffraction is observed at a harmonic structure. The diffraction angle is connected with the period of the harmonic structure through a relation similar to Eq. (33.36) for the first-order principal maximum at a diffraction grating having the same period.

We can now treat diffraction at a grating as diffraction at harmonic structures. The periodic function expressing the transmission coefficient of the grating with a period d can be expanded into a Fourier series in harmonic functions whose periods are equal to $d, d/2, d/3, \ldots$. The diffraction of a wave of wavelength λ, incident on the grating, is reduced to the diffraction at the harmonic structures constituting the grating, obtained as a result of the expansion of the transmission coefficient of the grating into a Fourier series. The diffraction by the harmonic component of the structure of length d/m results in two diffraction maxima whose conditions are given, in accordance with (33.64), by

$$(d/m)\sin\varphi_m = \lambda, \quad (d/m)\sin\varphi_{-m} = -\lambda \qquad (33.64b)$$

or

$$\underline{\overline{d\sin\varphi = m\lambda \;\; (m = \pm 1, \pm 2, \ldots),}} \qquad (33.65)$$

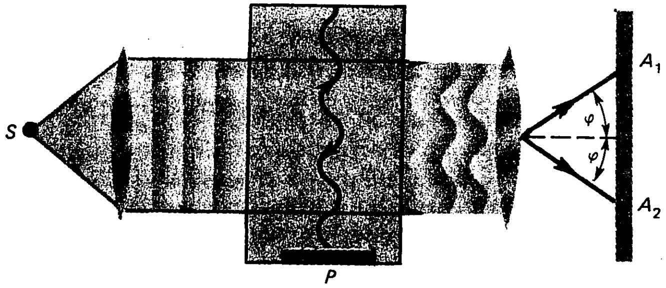


Fig. 177. Experimental device for observing diffraction by ultrasonic waves.

which is the same as (33.36). Thus, the position of the principal maxima of the grating is completely determined. In principle, such a method can be also used to describe a change in the intensity by considering the values of the coefficients in the Fourier series.

The process of analyzing the diffraction at an arbitrary periodic structure is now clear. We must present its characteristics through a Fourier series and consider the first-order diffraction described by the individual terms in the Fourier series. These first-order diffraction patterns together form the entire diffraction pattern on the periodic structure. It is clear that, in principle, diffraction at an aperiodic structure can be considered similarly, except that the Fourier series must be replaced by the Fourier integral.

DIFFRACTION BY ULTRASONIC WAVES. Ultrasonic vibrations have a frequency of the order of 10^8 Hz. In a liquid, the velocity of sound is $v \simeq 10^3$ m/s, and hence the wavelength of an ultrasonic wave is $d = v/\nu = 10^{-5}\ \mathrm{m} = 10\ \mu\mathrm{m}$. Condensations and rarefactions in an ultrasonic wave propagating in a liquid create a harmonic phase grating. The harmonic phase modulation leads to a diffraction similar to that observed for the harmonic amplitude modulation. Hence we must observe a first-order diffraction which can be conveniently reproduced by using the ultrasonic device shown in Fig. 177. Piezoelectric transducer P generates ultrasonic waves which cause the diffraction of waves emitted by source S. In accordance with (33.64a), we observe two first-order maxima and a central maximum.

COMPARISON OF THE CHARACTERISTICS OF SPECTRAL INSTRUMENTS. One of the main applications of interferometers and diffraction gratings involves the spectral analysis of radiation. In other words, these instruments are used as spectral devices. The expedience in using a particular type of instrument depends on the specific conditions of the problem.

High resolving powers can be attained in the Fabry-Perot and Michelson interferometers ($\sim 10^6$) as well as in diffraction gratings ($\sim 10^5$) and other types of interferometers. However, different factors are responsible for the high resolving power in these instruments. In the Fabry-Perot and Michelson interferometers, this is made possible by the high order of interference (about 10^5) for a comparatively small number of interfering rays (a few dozens in the Fabry-Perot interferometer and only two rays in the Michelson interferometer), while in a diffraction grating the large number (about 10^5) of interfering rays for a low order of interference (of the order of unity) is responsible for the high resolving power. Consequently, the dispersion region is very small in the Fabry-Perot interferometer (about 10^{-2} nm) and the Michelson interferometer (about 10^{-3} nm) and is very large in a diffraction grating (about 10^3 nm). Therefore, if the radiation under consideration has a large dispersion region, and a high-resolution instrument with a small dispersion region is required to analyze it, we have to use combinations of various spectral instruments. This yields simultaneously a high resolving power and a large dispersion region.

Spectral instruments must ensure the possibility of operating with a low-intensity radiation. In this respect, the Fabry-Perot interferometer has considerable advantages over the diffraction grating, especially if a photoelectric detecting system is used in the scanning Fabry-Perot interferometer. In Fourier spectroscopy, the resolving power is determined by the maximum path difference, which can be controlled by moving mirrors and attains large values.

Example 33.1. Calculate Fraunhofer's diffraction at a slit having a transmission coefficient $\tau(y') = \cos(\pi y'/b)$. The origin of coordinates is located at the centre of the slit. The width of the slit is b and its diffraction angle is φ. We have

$$A_\varphi = B \int_{-b/2}^{b/2} \cos(\pi y'/b)\, \mathrm{e}^{-ik\sin\varphi y'}\, \mathrm{d}y' = (B/2) \int_{-b/2}^{b/2} \mathrm{e}^{i(\pi/b - k\sin\varphi)y'}\, \mathrm{d}y'$$

$$+ (B/2) \int_{-b/2}^{b/2} \mathrm{e}^{-i(\pi/b + k\sin\varphi)y'}\, \mathrm{d}y' = -\frac{2\pi Bb\cos(kb\sin\varphi/2)}{(kb\sin\varphi)^2 - \pi^2}. \quad (33.66)$$

For $\varphi \to 0$, the amplitude is given by

$$A_0 = 2bB/\pi. \quad (33.67)$$

Hence we obtain for the intensity distribution

$$I_\varphi = \frac{I_0}{[1 - 4b^2\sin^2\varphi/\lambda^2]^2}\cos^2(\pi b\sin\varphi/\lambda), \quad (33.68)$$

where I_0 is the intensity in the direction $\varphi = 0$.

SEC. 34. FRESNEL DIFFRACTION

Diffraction in the near region is considered and a quantitative description of Cornu's spiral is given.

FRESNEL'S DIFFRACTION REGION. The Fresnel diffraction region lies in the vicinity of the object at which the diffraction takes place and extends to distances from where diffraction can be considered as Fraunhofer's diffraction [see (33.7)]. However, it is worth noting that in a strict sense, Fraunhofer's diffraction takes place only at infinity.

In order to consider Fresnel's diffraction, we must use formula (32.37) in which the factor in front of the integral can be omitted in order to simplify notation since it does not affect the relative intensity distribution in the diffraction pattern.

DIFFRACTION AT A RECTANGULAR APERTURE. The position of the coordinate system and the size of the rectangular aperture are shown in Fig. 159. Assuming that a plane wave with amplitude A_0 is incident on the aperture, we put $\Psi = A_0/(ab)$ in (32.37) and write this formula as

$$A(x,y) = [A_0/(ab)\int_{-a/2}^{a/2} \exp[ik(x-x')^2/(2l)]\,\mathrm{d}x \times \int_{-b/2}^{b/2} \exp[ik(y-y')^2/(2l)]\,\mathrm{d}y'. \tag{34.1}$$

The first of the integrals on the right-hand side can be evaluated as follows:

$$Q = \int_{-a/2}^{a/2} \exp\left[\frac{ik(x-x')^2}{2l}\right]\mathrm{d}x'. \tag{34.2}$$

Going over to the new integration variable

$$\xi = (x-x')[k/(\pi l)]^{1/2}, \quad \mathrm{d}\xi = -\,\mathrm{d}x'\,[k/(\pi l)]^{1/2}, \tag{34.3}$$

we obtain

$$Q = -(\pi l/k)^{1/2}\int_{(x+a/2)\sqrt{k/(\pi l)}}^{(x-a/2)\sqrt{k/(\pi l)}} \mathrm{e}^{i\pi\xi^2/2}\,\mathrm{d}\xi$$
$$= (\pi l/k)^{1/2}\left[\int_0^{(x+a/2)\sqrt{k/(\pi l)}} \mathrm{e}^{i\pi\xi^2/2}\,\mathrm{d}\xi - \int_0^{(x-a/2)\sqrt{k/(\pi l)}} \mathrm{e}^{i\pi\xi^2/2}\,\mathrm{d}\xi\right]. \tag{34.4}$$

FRESNEL INTEGRALS. Let us evaluate

$$\int_0^u \mathrm{e}^{i\pi\xi^2/2}\,\mathrm{d}\xi = \int_0^u \cos(\pi\xi^2/2)\,\mathrm{d}\xi + i\int_0^u \sin(\pi\xi^2/2)\,\mathrm{d}\xi. \tag{34.5}$$

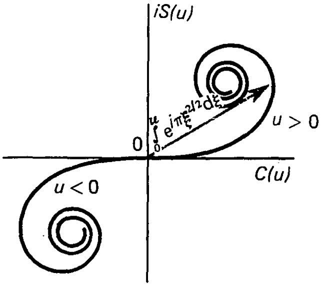


Fig. 178. Construction of Cornu's spiral with the help of Fresnel's integrals.

The integrals

$$C(u)=\int_0^u \cos(\pi\xi^2/2)\,d\xi,\quad S(u)=\int_0^u \sin(\pi\xi^2/2)\,d\xi \tag{34.6}$$

are called Fresnel's integrals and are determined by numerical calculations. There exist tables of these integrals which can be graphically presented in the form of Cornu's spiral (see Fig. 151). A more precise construction of this spiral with the help of (34.6) is shown in Fig. 178. It should be noted that $C(\infty)=S(\infty)=0.5$, $C(-\infty)=S(-\infty)=-0.5$.

CORNU'S SPIRAL. From definition (34.5) and the rules for constructing Cornu's spiral, it follows that the complex number

$$\int_0^u e^{i\pi\xi^2/2}\,d\xi \tag{34.7}$$

is represented by a complex vector starting at the origin of coordinates and terminating at a point on Cornu's spiral corresponding to the value u (Fig. 178). This substantiates the method of diffraction analysis with the help of Cornu's spiral (see Sec. 31). It can be used to find integrals (34.1) and to analyze quantitatively the intensity distribution in the diffraction pattern.

PROBLEMS

6.1. The wavelengths of the sodium doublet lines are $\lambda_1 = 589$ nm and $\lambda_2 = 589.6$ nm. What is the minimum number of lines in a diffraction grating required to resolve this doublet (a) in the second order and (b) in the first order?

6.2. A monochromatic wave is made to pass through a glass plate covered with lycopodium powder. With the help of a convex lens of focal length 1 m, an interference pattern is observed on the screen behind the plate. The central dark spot has a radius of 3 cm. The wavelength of the radiation is 600 nm. Find the radius of the lycopodium powder particles.

6.3. A diffraction grating has 200 lines per millimetre. At what angle should radiation be incident on the grating to observe a second-order diffraction maximum for the wavelength $\lambda = 0.5$ m in a direction perpendicuar to the grating?

6.4. Light is normally incident on a grating. Find the ratio of the intensities I_m and I_0 of the m-th order principal maximum and the central maximum.

6.5. A parallel beam of monochromatic light is incident on a wire of diameter 1 mm stretched in a direction perpendicular to the direction of light propagation. Diffraction fringes separated by 0.5 mm are observed on a screen located at a distance of 1 m from the wire and perpendicular to the direction of light propagation. Find the wavelength of the radiation incident on the wire.

6.6. Under what condition does the m-th principal maximum for a diffraction grating with a period d and slit width b vanish?

6.7. A disc of diameter 0.5 cm having irregularities of the order of 100 μm is placed at a distance of 1 m from a monochromatic point source of light emitting radiation with a wavelength of 0.5 μm. The point source is located on a line perpendicular to the plane of the disc and passing through its centre. Assuming that the Poisson spot is visible only so long as the irregularities of the disc edges do not cover more than 1/4 of the width of the corresponding Fresnel zone, find the minimum distance at which the Poisson spot is visible.

6.8. A monochromatic point source of light ($\lambda = 0.5$ μm) is located at a distance of 0.5 m from a plane screen with a circular aperture of diameter 0.5 mm on the perpendicular to the plane of the screen, passing through the centre of the aperture. What is the distance between the aperture and the point with the maximum illuminance located on the perpendicular?

6.9. Find the condition under which all the even maxima vanish in the diffraction pattern formed by a grating.

6.10. The surface of the Moon is irradiated by a laser beam whose diameter at the outlet from the laser is 0.5 cm. The beam divergence is purely diffractional, and the influence of the atmosphere can be neglected. Find the diameter of the laser beam as it reaches the surface of the Moon, assuming that the wavelength of the laser radiation is 632.8 nm and the distance between the Earth and the Moon is 368×10^3 km.

6.11. The entrance pupil of the eye has a radius of 1.5 mm. At what distance does a luminous disc of diameter 1 cm emitting radiation of wavelength $\lambda = 0.55$ μm produce a coherent field on the entrance pupil?

6.12. A Michelson echelon consists of twenty 1 mm-thick plane-parallel plates with a refractive index $n = 1.5$. The height of the steps is 0.5 cm. Find the angular width of the diffraction maxima and the order of interference. The wavelength of the light incident on the echelon is $\lambda = 0.55$ μm.

6.13. Find the angular dispersion of a diffraction grating having 4×10^3 lines per centimetre in the second-order spectrum.

6.14. What is the angular resolution of the human eye (pupil diameter 4 mm) and a telescope (diameter of the entrance pupil 10 cm) for light of wavelength 0.6 μm?

ANSWERS

6.1. 491; 982. 6.2. 12 μm. 6.3. 23°30′. 6.4. $I_m/I_0 = [\sin(mb\pi/d)/(mb\pi/d)]^2$. 6.5. 0.5 μm. 6.6. $m = nd/b$ (n is an integer). 6.7. 0.67 m. 6.8. 0.17 m. 6.9. $d = 2b$. 6.10. 118.58 km. 6.11. 110 m. 6.12. 0.27×10^{-4} rad; 4.5×10^3. 6.13. 0.8×10^5 rad/m = 0.08 rad/μm. 6.14. $1.83 \cdot 10^{-4}$ rad = 38″; 7.32×10^{-6} rad = 1.5″.

Chapter 7

Basic Concepts of Fourier Optics

Basic idea:
An object is considered as a source of secondary waves. The diffraction pattern from this object is then described by the methods of the diffraction theory.

SEC. 35. LENS AS AN ELEMENT ACCOMPLISHING A FOURIER TRANSFORMATION

It is shown that a lens accomplishes a Fourier transformation of the amplitude distribution between the first and second focal planes of the lens with an error not exceeding a phase multiplier which is insignificant for the intensity distribution.

PHASE TRANSFORMATION BY A THIN LENS. A thin lens is one in which the coordinates (x', y') of the incoming and outgoing rays can be assumed to be the same. The refractive index of the material of the lens is denoted by n (Fig. 179).

The planes perpendicular to the Z-axis and tangential to the left and right surfaces of the lens are designated as left and right planes respectively. We denote the coordinates of points in these planes by (x', y') since according to the definitions of a thin lens, a ray of light always joins these points. The length of the ray characterized by the points (x', y') in the lens is denoted by $\Delta(x', y')$, while the maximum thickness of the lens is Δ_0 (Fig. 179).

The total variation in the phase of a wave upon passing from the left to the right plane is

$$\begin{aligned}\delta(x',y') &= kn\Delta(x',y') + k[\Delta_0 - \Delta(x', y')]\\ &= k\Delta_0 + k(n-1)\Delta(x',y'),\end{aligned} \tag{35.1}$$

where $\Delta(x',y')$ is a function of the lens thickness. The quantity

$$\tau(x',y') = e^{i\delta(x',y')} = e^{ik\Delta_0}e^{ik(n-1)\Delta(x',y')} \tag{35.2}$$

is called the transmission coefficient of the lens. It is a purely phase quantity for a lens since the light absorption in this case is negligibly small.

In geometrical optics, a ray was taken as the basic concept. The effect of various factors on a ray of light was determined in

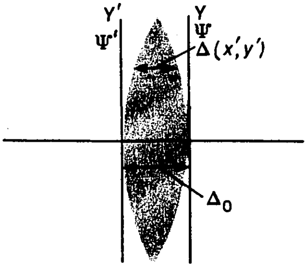


Fig. 179. To the calculation of the phase transformation by a lens.

the problems of geometrical optics. This effect was reduced to a change in the direction of a ray under the action of these factors. For example, as a result of refraction at the surfaces of a double convex lens, a ray of light deviates from its original direction towards the axis of the lens. The concept of a ray was connected with the wave nature of light through its definition: a ray is a line the tangent to which at any point coincides with the normal to the wave front at this point. In subsequent analysis, there is no need to refer to the wave nature of light.

The wave concept is the underlying principle of the Fourier optics. The basic objective is to determine the effect of various factors on the wave front, phase and amplitude of a wave. Such an approach is of a more general nature than in the geometrical optics since a knowledge of the wave front allows the determination of not only the corresponding rays, but also other properties of the radiation, including the distribution of amplitudes.

To begin with, let us consider the passage of an electromagnetic wave through a thin lens.

The amplitude $\Psi'(x', y')$ of a wave entering a lens at the left plane and the amplitude $\Psi(x', y')$ of the wave leaving the lens through the right plane are connected through the relation

$$\Psi(x', y') = \tau(x', y')\Psi'(x', y'). \tag{35.3}$$

EVALUATION OF THE THICKNESS FUNCTION. While calculating the thickness function of a lens, the radius of curvature of the convex surface in the direction of ray propagation is assumed to be positive, and the radius of curvature of the concave surface is assumed to be negative (Fig. 180). It is assumed that the ray travels from left to right. This gives

!

A lens transforms the phase of a plane wave so that the wave becomes a convergent or divergent spherical wave.
To within insignificant scale and phase factors, the amplitude distribution in the focal plane of the lens is the Fourier transform of the amplitude distribution at the entrance of the lens.

$$\begin{gathered}\Delta = \Delta_1 + \Delta_2, \quad \Delta_0 = \Delta_{01} + \Delta_{02}, \quad \rho^2 = x'^2 + y'^2, \\ \Delta_1 = \Delta_{01} - (R_1 - \sqrt{R_1^2 - \rho^2}), \\ \Delta_2 = \Delta_{02} - (-R_2 - \sqrt{R_2^2 - \rho^2}),\end{gathered} \tag{35.4}$$

whence

$$\Delta = \Delta_1 + \Delta_2 = \Delta_0 - R_1(1 - \sqrt{1 - \rho^2/R_1^2}) + R_2(1 - \sqrt{1 - \rho^2/R_2^2}), \tag{35.5}$$

where we have taken into account the fact that in the last term on the right-hand side of (35.5), the square root of R_2^2 is negative, i.e. $\sqrt{R_2^2} = -R_2$. Expanding the roots into a series in $(\rho/R)^2$ and confining the expansion to terms of the order of

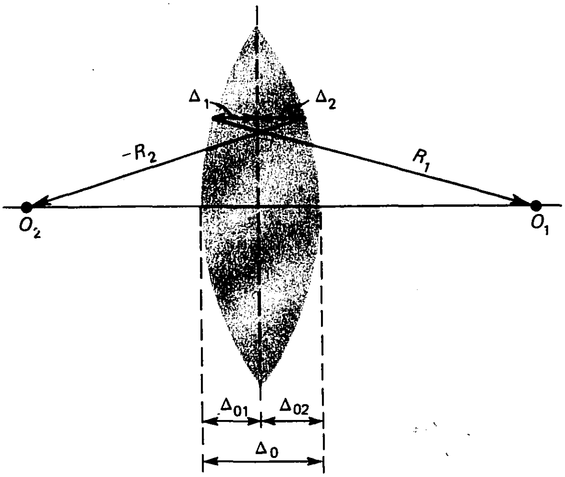


Fig. 180. To the calculation of the lens thickness function.

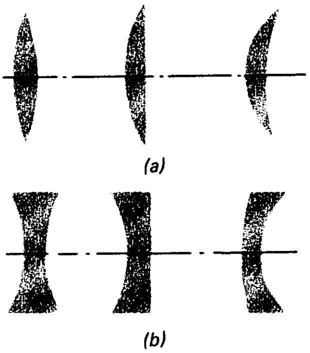


Fig. 181. Lenses with positive ($f > 0$) (*a*) and negative ($f < 0$) (*b*) focal lengths.

$(\rho/R)^2$, we obtain

$$\Delta = \Delta_0 - \frac{x'^2 + y'^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{35.6}$$

On the basis of (23.32), we can write

$$1/f = (n-1)(1/R_1 - 1/R_2) \tag{35.7}$$

and present (35.6) in the form

$$\Delta = \Delta_0 - (x'^2 + y'^2)/[2f(n-1)], \tag{35.8}$$

where f is the focal length of the lens. Substituting (35.8) into (35.1), we obtain

$$\delta(x', y') = kn\Delta_0 - k(x'^2 + y'^2)/(2f). \tag{35.9}$$

Consequently, (35.2) assumes the form

$$\tau(x', y') = \exp(ikn\Delta_0)\exp\{-ik(x'^2 + y'^2)/(2f)\}. \tag{35.10}$$

TYPES OF LENSES. Lenses with a positive focal length ($f > 0$) are shown in Fig. 181*a*, and with a negative focal length ($f < 0$) in Fig. 181*b*. The first factor on the right-hand side of (35.10) does not lead to a distortion of the shape of a plane wave passing through the lens, while the second factor transforms a plane wave into a spherical wave which may converge or diverge, depending on the sign of the exponent. If the exponent is negative (i.e. $f > 0$), the peripheral parts of the wave passing through the lens undergo a smaller phase lag than the middle of the wave. As a result, a converging spherical wave is formed at the exit from the lens (Fig. 182). For $f < 0$, a diverging spherical wave leaves the lens (Fig. 183). Such

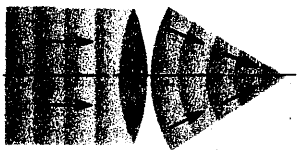


Fig. 182. A plane wave emerging from a lens (with $f > 0$) as a converging spherical wave.

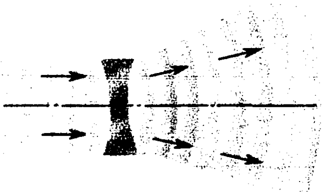


Fig. 183. A plane wave emerging from a lens (with $f < 0$) as a diverging spherical wave.

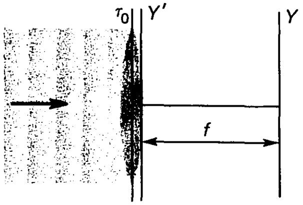


Fig. 184. Formation of the image of an object placed in front of a lens.

a wave is spherical only in paraxial approximation. In this case, aberrations are not taken into account.

LENS AS AN ELEMENT ACCOMPLISHING A FOURIER TRANSFORMATION. We place an object with an amplitude transmission coefficient $\tau_0(x', y')$ in front of a lens (Fig. 184) and direct a monochromatic plane wave towards it. A light field $A_0\tau_0(x', y') = \Psi'(x', y')$ is formed in the first plane in front of the lens, A_0 being the amplitude of the plane wave. The finite size of the lens aperture is taken into account through the pupil function defined by the condition

$$P(x', y') = \begin{cases} 1 & \text{within the lens aperture} \\ 0 & \text{beyond the lens aperture.} \end{cases} \tag{35.11}$$

Combining (35.3) and (35.10), we can present the light field formed at the exit from the lens through the function

$$\Psi(x', y') = \exp(ikn\Delta)\,\Psi'(x', y')\,P(x', y')$$
$$\times \exp[-ik(x'^2 + y'^2)^{\tau}/(2f)\,i. \tag{35.12}$$

For $z_0 = f$, the diffraction pattern in the focal plane of the lens can be defined with the help of (32.37) as follows:

$$\Phi(x, y) = \frac{k}{2\pi i}\,\frac{\exp[i(kf + kn\Delta_0)]}{f} \iint\limits_{-\infty}^{\infty} \Psi'(x', y')\,P(x', y')$$
$$\times \exp\left[-\frac{ik(x'^2 + y'^2)}{2f}\right] \exp\left\{\frac{ik[(x - x')^2 + (y - y')^2]}{2f}\right\} \mathrm{d}x'\mathrm{d}y', \tag{35.13}$$

where the integration is carried out from $-\infty$ to $+\infty$ since the pupil function P in the integral automatically ensures that integration is carried out over the aperture area.

Taking (33.1) into account, we can present formula (35.13) as

$$\Phi(x, y) = \frac{k}{2\pi i}\,\frac{\exp[i(kf + kn\Delta_0)]}{f} \exp\left[\frac{ik(x^2 + y^2)}{2f}\right]$$
$$\times \iint\limits_{-\infty}^{\infty} \Psi'(x', y')\,P(x', y') \exp\left[-\frac{ik(xx' + yy')}{f}\right] \mathrm{d}x'\mathrm{d}y'. \tag{35.14}$$

It can be seen that the amplitude distribution in the focal plane is the Fourier transform of the amplitude distribution at the lens entrance with an error not exceeding certain phase multipliers and scale factors that are insignificant from the point of view of intensity distribution. In other words, a lens is an element that brings about a Fourier transformation.

?

How is the presence of a diaphragm at the lens entrance taken into account?

Why is the Fourier transform of the amplitude distribution at the lens entrance formed just in the focal plane of the lens and not in any other plane?

SEC. 36. DIFFRACTION FORMATION OF IMAGES BY A LENS

The process of transformation of a diffraction pattern into the image of an object is described. The limiting resolving power of optical instruments is discussed.

FOURIER TRANSFORMATION OF AMPLITUDES BETWEEN THE FOCAL PLANES OF A LENS. The ideas presented in Sec. 35 show that in the process of wave propagation, the amplitude distribution in a plane perpendicular to the direction of wave propagation varies from one plane to another. By a successive application of the formulas describing these variations, we can obtain the transformation relations for the amplitude distribution between any two planes. The intensity distribution in these planes can also be determined. The relation between the amplitude distributions in the general case is quite complicated, and the intensity distributions are quite different from one another. Under certain conditions, however, the relation between amplitude distributions is quite simple and its main part can be reduced to a Fourier transform. Naturally, the simplest cases will be considered in the first place. After this, we shall consider the conditions under which the intensity distributions in two planes are fairly similar. Then we can speak of the diffraction image formation since the entire analysis is based on the wave concept and no reference is made to the rays. We place a plane object with an amplitude transmission coefficient $\tau_0(x_0, y_0)$ in front of a lens at a distance L (Fig. 185) and direct a plane monochromatic wave at it. In the second focal plane of the lens, a light field

$$\Psi''(x_0, y_0) = A_0\tau_0(x_0, y_0) \tag{36.1}$$

is produced, A_0 being the amplitude of the plane wave falling on the object. This light field produces a diffraction pattern $\Psi'(x', y')$ in the first focal plane of the lens. With the help of (32.37), we can present this diffraction pattern in the form

$$\Psi'(x', y') = \frac{k}{2\pi i}\frac{e^{ikL}}{L}\iint \Psi''(x_0, y_0) \times \exp\left\{\frac{ik[(x' - x_0)^2 + (y' - y_0)^2]}{2L}\right\} dx_0 dy_0, \tag{36.2}$$

where the integration with respect to x_0 and y_0 is carried out between $-\infty$ and ∞, and the finite size of the object is taken into consideration by putting the transmission coefficient $\tau_0(x_0, y_0)$ equal to zero outside the object. If the contribution to the diffraction from the wave front outside the object has to be taken into consideration, $\tau_0(x_0, y_0)$ should be put equal to unity for these parts of the wave front.

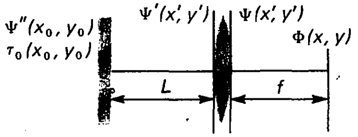


Fig. 185. Formation of the image of an object placed in front of a lens at a distance L.

The light field at the exit from the lens is given by (13.12), and in the focal plane by formula (35.13) which becomes

$$\Phi(x,y) = B \iint_{-\infty}^{\infty} \exp\left\{\frac{ik[(x-x')^2+(y-y')^2]}{2f}\right\} dx'dy'P(x',y')$$
$$\times \exp\left[-\frac{ik(x'^2+y'^2)}{2f}\right] \iint_{-\infty}^{\infty} \Psi''(x_0,y_0)$$
$$\times \exp\left\{\frac{ik[(x'-x_0)^2+(y'-y_0)^2]}{2L}\right\} dx_0dy_0, \tag{36.3}$$

where B stands for all the constants in front of the integrals appearing in expressions (36.3).

Expanding the squared coordinate differences in the exponents in the integrand, we can write (36.3) in the form

$$\Phi(x,y) = Be^{ik(x^2+y^2)/(2f)} \iiiint_{-\infty}^{\infty} P(x',y')\Psi''(x_0y_0)\,e^{-ik(xx'+yy')/f}$$
$$\times e^{ik(x'^2+y'^2)/(2L)} e^{ik(x_0^2+y_0^2)/(2L)} e^{-ik(x'x_0+y'y_0)/L} dx'dy'dx_0dy_0. \tag{36.4}$$

The pupil function P affects the diffraction pattern in a definite way. If the pupil is not very small, then $P=1$ for all x', y', since the exponential functions depending on these coordinates oscillate strongly outside the pupil and the contribution to the integral from the region outside the pupil becomes negligibly small. The condition for integration with respect to x' and y' in (36.4) can then be satisfied by using the following formula from the tables of integrals:

$$\int_{-\infty}^{\infty} \exp[i(a\xi^2+2b\xi)]\,d\xi = (1+i)\sqrt{\pi/(2a)}\exp(-ib^2/a).$$

As a result of integration, (36.4) assumes the form

$$\Phi(x,y) = B_1 \exp\left[\frac{ik}{2f}\left(1-\frac{L}{f}\right)(x^2+y^2)\right]$$
$$\times \iint_{-\infty}^{\infty} \Psi''(x_0,y_0)\exp\left[-\frac{ik(xx_0+yy_0)}{f}\right]dx_0dy_0, \tag{36.5}$$

where B_1 combines all constant multipliers. This means that **the amplitude distribution in the focal plane of a lens is the Fourier transform of the amplitude distribution for the light field on the surface of the object with an error not exceeding certain insignificant phase multiplier and scale factor.** If the object is located in the first focal plane of the lens ($L=f$), the phase multiplier vanishes. This means that the amplitude

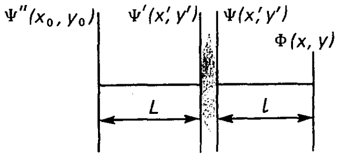


Fig. 186. To the analysis of the image formation by a lens.

distributions in the focal planes of the lens are connected through the Fourier transform without any phase distortions, i.e. the lens brings about a phase transformation between the amplitude distributions of a light field at its focal planes.

IMAGE FORMATION BY A LENS. If the amplitude distribution is to be determined not in the second focal plane of the lens but in a plane at a distance l from the lens (Fig. 186), we make use of formulas (36.1)-(36.3). In formula (36.3), the first exponent in the integrand now contains, instead of the distance f between the lens and the second focal plane, the distance l between the lens and the plane in which the amplitude distribution is to be determined. Then [see (36.4)]

$$\Phi(x, y) = B_2 \int_{-\infty}^{\infty} G(x, y, x_0, y_0) \Psi''(x_0, y_0) \, dx_0 dy_0, \tag{36.6}$$

where

$$G(x, y, x_0, y_0) = \exp\left[\frac{ik(x^2 + y^2)}{2l}\right] \exp\left[\frac{ik(x_0^2 + y_0^2)}{2L}\right]$$
$$\times \iint_{-\infty}^{\infty} P(x', y') \exp\left[\frac{ik}{2}\left(\frac{1}{L} + \frac{1}{l} - \frac{1}{f}\right)(x'^2 + y'^2)\right]$$
$$\times \exp\left\{-ik\left[\left(\frac{x_0}{L} + \frac{x}{l}\right)x' + \left(\frac{y_0}{L} + \frac{y}{l}\right)y'\right\} dx'dy'. \tag{36.7}$$

If the intensity distribution $|\Phi(x, y)|^2$ resembles the distribution of the intensities $|\Psi''(x_0, y_0)|^2$ in the plane of the object, the diffraction pattern $|\Phi(x, y)|^2$ is referred to as the image of the object. The image may be larger or smaller than the object, erect or inverted. The concept of the image is introduced in geometrical optics by attributing to each point on the object a point on the image in accordance with the rules considered in Sec. 25. Hence we can compute (36.7) and (36.6) in the geometrical optics approximation:

$$\lambda \to 0, \quad k \to \infty. \tag{36.8}$$

Let us analyze the integral appearing in (36.7). We choose the distances L and l in such a way that the term with the exponent containing the sum $x'^2 + y'^2$ becomes equal to unity. For this purpose, we must put

$$1/L + 1/l - 1/f = 0. \tag{36.9}$$

Then

$$Q = \iint_{-\infty}^{\infty} P(x', y') \exp\left\{-ik\left[\left(\frac{x_0}{L} + \frac{x}{l}\right)x' + \left(\frac{y_0}{L} + \frac{y}{l}\right)y'\right]\right\} dx'dy'. \tag{36.10}$$

!

The image is the term applied to a diffraction pattern in which the light intensity reproduces, to a sufficiently high degree of accuracy, the intensity distribution of the original object.

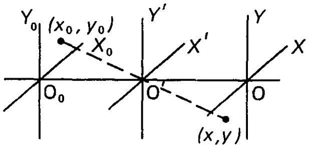


Fig. 187. To the derivation of the condition for image formation by a lens.

Denoting

$M = l/L,$

we can present (36.10) in the form

$$Q = \iint_{-\infty}^{\infty} P(x', y') \exp\left\{ -\frac{ik}{l}[(x + Mx_0)x' + (y + My_0)y'] \right\} \mathrm{d}x' \mathrm{d}y'. \tag{36.12a}$$

Let us evaluate this integral in the geometrical optics approximation when $\lambda \to 0$ and $k \to \infty$. For this purpose, we go over to new integration variables in (36.12a): $\xi = (k/l)x'$, $\eta = (k/l)y'$. This gives

$$Q = \left(\frac{l}{k}\right)^2 \iint_{-\infty}^{\infty} P[(l/k)\xi, (l/k)\eta] \exp[-i(x + Mx_0)\xi - i(y + My_0)\eta] \,\mathrm{d}\xi \mathrm{d}\eta. \tag{36.12b}$$

As $k \to \infty$, the arguments of P for any finite values of ξ and η tend to zero, and hence $P \to 1$. For $P = 1$, integral (36.12b) can be presented through a delta-function:

$$Q = (l/k)^2 \iint_{-\infty}^{\infty} \exp[-i(x + Mx_0)\xi - i(y + My_0)\eta] \,\mathrm{d}\xi \,\mathrm{d}\eta = 4\pi^2 (l/k)^2 \delta(x + Mx_0) \delta(y + My_0). \tag{36.12c}$$

Hence in the geometrical optics limit (36.7) assumes the form

$$G(x, y, x_0, y_0) = 4\pi^2 (l/k)^2 \exp\left[\frac{ik(x^2 + y^2)}{2l}\right] \times \exp\left[\frac{ik(x_0^2 + y_0^2)}{2L}\right] \delta(x + Mx_0) \delta(y + My_0). \tag{36.13}$$

Substituting expression (36.13) into (36.6) and integrating with respect to x_0 and y_0, we get

$$\Phi(x, y) = B_4 \exp\left[\frac{ik(x^2 + y^2)}{2l}\right] \exp\left[\frac{ik(x^2 + y^2)}{2LM^2}\right] \times \Psi''\left(-\frac{x}{M}, -\frac{y}{M}\right), \tag{36.14}$$

where all the constants that are not significant for the distribution of the amplitudes $\Phi(x, y)$ are included in B_4.

From (36.14) we obtain

$$|\Phi(x, y)|^2 = |B_4|^2 |\Psi''(-x/M, -y/M)|^2. \tag{36.15}$$

!

Theoretically an exact image of a plane object can be obtained only with the help of unrestricted light beams. In the presence of a diaphragm, each point of the object is represented by the diffraction pattern of the diaphragm. This restricts the resolving power of optical instruments.

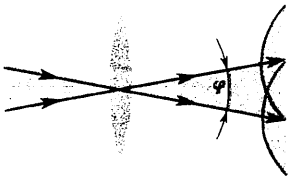


Fig. 188. To the calculation of the limiting resolving power of a telescope objective.

This formula shows that the intensity distribution $|\Phi(x, y)|^2$ observed in the diffraction pattern under geometrical optics approximation is an inverted image of the intensity distribution $|\Psi''|^2$ with a magnification M (Fig. 187). Thus, the image is formed when condition (36.9) is satisfied. In geometrical optics, this condition is written in the form

$$1/L + 1/l = 1/f, \tag{36.16}$$

where L is the distance between the lens and the object and l is the distance between the lens and the image. This means that in the geometrical optics approximation, formula (36.16) describes the formation of geometrical images in accordance with the laws of geometrical optics.

However, the image structure becomes complicated if we consider finite wavelengths and take into account the finite size of the lens aperture. Each point on the object is no longer represented by a point on the image. A comparison of (36.6), in which G is taken in the form (36.13), with (33.12a) shows that the image of the object is distorted by Fraunhofer's diffraction at the lens aperture. This is taken into account in formula (36.12a) through the pupil function $P(x', y')$. Each point on the object produces in the image plane a Fraunhofer diffraction pattern in the lens aperture, i. e. the point is represented by an image in the form of a bright spot. In the case of a circular aperture, the radius of the central bright spot can be calculated with the help of (33.27).

LIMITING RESOLVING POWER OF OPTICAL INSTRUMENTS. The resolving power of optical instruments is limited by the Fraunhofer diffraction at their entrance aperture since in this case each point of the object is represented by a diffraction pattern whose centre lies at the point corresponding to the ideal geometrical image. The diffraction pattern consists of a central bright spot surrounded by diffraction rings. To a fairly high degree of accuracy, the central bright spot of the diffraction pattern can be taken as the image of the point responsible for this pattern. Two points on an object will be resolved if the bright spots serving as their images can be separately identified. For the criterion of resolution, we take the Rayleigh criterion.

Let us determine the limiting resolving power of the objective of a telescope. The object is located at infinity and the image is formed in the focal plane of the objective with focal length f (Fig. 188). It can be concluded on the basis of (33.27) that the radius of the central bright spot is

$$r_0 = 0.61 f\lambda/R, \tag{36.17}$$

?

What is the difference in the images of an object, obtained by the phase-contrast method, if additional phase shifts of $\pi/2$ and $3\pi/2$ are introduced in the light passing by the object? Why cannot the "dark-field method" be replaced by the "bright-field method", i. e. why cannot an object be registered from absorbed radiation?

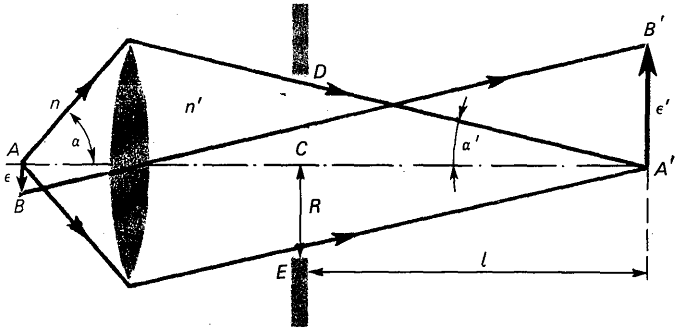


Fig. 189. To the calculation of the limiting resolving power of a microscope.

where R is the radius of the telescope objective. Two closely situated stars having an angular distance θ form two bright diffraction spots having a radius r_0 each and a distance $\Delta = \theta_0 f$ between their centres. According to Rayleigh's criterion, these two stars are considered to have been resolved if the centre of one spot falls on the edge of the other, i.e. when $\Delta = r_0$, and hence

$$\theta_0 = 0.61\lambda/R. \tag{36.18}$$

The resolving power Υ of the objective is the quantity reciprocal to θ_0:

$$\Upsilon = 1/\theta_0 = R/(0.61\lambda). \tag{36.19}$$

In order to increase the resolving power of an objective, we must increase its aperture.

If the angular distance between two points is less than θ_0, they are treated as a single point, and hence the details of an object with smaller angular size cannot be distinguished.

Formula (36.19) for the resolving power of an objective is obtained on the basis of relation (33.27) which was introduced under the assumption that the amplitude and phase of the incident wave are constant at all points on the objective aperture (uniform aperture). Under this condition, the only way to increase the resolving power for a fixed wavelength is to increase the radius of the objective. However, the radius of the central diffraction maximum in the diffraction pattern formed by a circular aperture of a fixed radius can be decreased by a special choice of the distribution of amplitude and phase of the radiation in the objective plane, and this leads to a higher resolving power of the objective. However, the intensity of the central maximum is reduced in this case. Consequently, if a decrease in the brightness of the image is admissible, the resolving power of the objective can be increased (without

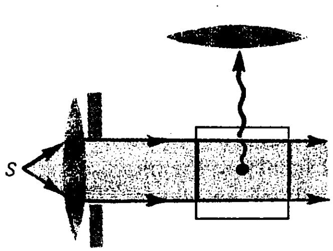


Fig. 190. Dark-field illumination method.

increasing its radius) through an appropriate phase and amplitude modulation of the light incident on the objective.

Similarly, we can find the limit of the resolving power of a microscope the path of the rays in which is shown in Fig. 189. In order to eliminate spherical aberration (see Sec. 24), Abbe's sine condition (24.17) must be satisfied. For the notation used in Fig. 189, this condition can be written in the form

$$\varepsilon n \sin\alpha = \varepsilon' n' \sin\alpha', \tag{36.20}$$

where n and n' are the refractive indices of the media in which the object and the image are located. Denoting by l the distance between the aperture plane and the image plane, and by R the radius of the entrance pupil, we find that the radius of the central bright spot in the diffraction pattern is

$$r_0 = l\varphi = 0.61\lambda l/R, \tag{36.21}$$

where $R = DE/2$.

The first stage of image formation in microscope is accomplished by the objective. The image obtained in this way is then magnified by the eyepiece. However, it is clear that the **resolving power is determined only by the first stage of image formation.** Hence Figure 189 shows the path of rays only in the microscope objective.

Points A and B can be distinguished from each other if, in accordance with Rayleigh's criterion, the distance between points A' and B' of their geometrical image is greater than or equal to r_0. Consequently, the maximum resolving power of the system (Fig. 189) is determined from the condition

$$\varepsilon' = r_0 = 0.61\lambda l/R. \tag{36.22}$$

From (36.20), we obtain

$$\varepsilon = \varepsilon' n' \sin\alpha'/(n\sin\alpha) = n'\lambda 0.61/(n\,sin\,\alpha), \tag{36.23}$$

where $R/d \approx \alpha' \approx \sin\alpha'$. **Formula (36.23) defines the maximum resolving power of a microscope.** Since $n' = 1$ for the medium where the image is usually formed, and the refractive index of the liquids used for oil immersion lenses (see Sec. 24) is $n \simeq 1.5$, while α is close to $\pi/2$, we can write (36.23) in the form

$$\varepsilon \simeq 0.4\lambda, \tag{36.24}$$

i. e.

$$\varepsilon \simeq 0.15 \text{ to } 0.3\ \mu\text{m} \tag{36.25}$$

for the visible region.

Finer details cannot be observed in the visible range.

DARK-FIELD ILLUMINATION METHOD. Parts of objects smaller than the values of ε determined by formula (36.25) cannot be distinguished in the optical region although

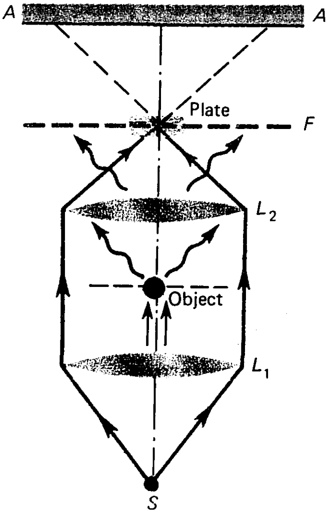


Fig. 191. Phase-contrast method.

minute particles of this size or even less can be detected. This is done with the help of dark-field illumination method (Fig. 190).

The particles under observation are illuminated by white light, from a source S in such a way that the direct light does not fall on the eye of the observer focussing the microscope at the point where the particle is likely to be present. The rays diffracted by the particle enter the microscope and are detected by the eye as luminous spots whose shape has got nothing to do with the shape of the particle. There also exist other methods in which direct rays are eliminated from the field of view by a special condenser, but the essence of the dark-field illumination method remains the same.

PHASE-CONTRAST METHOD. Thin transparent objects are very poor absorbers of light and do not produce any noticeable amplitude modulation of the transmitted light. As a result, they remain practically invisible. They can, however, be observed by modulating the phase of the light transmitted through them. The method in which phase modulation is employed for observing thin transparent objects is called phase-contrast method.

Light from a point source S (Fig. 191) is transformed into a beam of parallel rays with the help of a lens L_1. Most of the beam bypasses an object O, and only a small part passes through it. We take the amplitude of the light bypassing the object as unity and denote the amplitude of the light diffracted at the object as $\varepsilon \ll 1$, since a very small part of the light flux falls on the object. The wave passing through the object suffers a phase change δ, and hence its complex amplitude is equal to $\varepsilon e^{i\delta}$. A diffraction pattern is formed in the focal plane of lens L_2. This pattern is transformed into the image of object O in plane AA'.

Light bypassing the object does not undergo a phase change, and hence the total intensity in the image plane is given by

$$2I = |1 + \varepsilon e^{i\delta}|^2 = 1 + \varepsilon^2 + 2\varepsilon \cos\delta \approx (1 + \varepsilon)^2 - \varepsilon\delta^2, \quad (36.26)$$

where the first δ-dependent term in the expansion of $\cos\delta$ into a series in δ has the order of δ^2, and is therefore extremely small. Hence the phase modulation does not cause any observable intensity modulation.

However, conditions can be created under which the intensity modulation sharply increases and becomes proportional to the phase modulation. For this purpose, the light bypassing the object is made to pass through a plate which changes its phase by $\pi/2$, and the complex amplitude then becomes equal to $e^{i\pi}/2 = i$. As before, the amplitude of the light passing through the object is $\varepsilon e^{i\delta}$. In the image plane, the intensity of the interference pattern is given by

$$2I = |i + \varepsilon e^{i\delta}|^2 = 1 + \varepsilon^2 + 2\varepsilon \sin\delta \approx 1 + \varepsilon^2 + 2\varepsilon\delta. \qquad (36.27)$$

This means that the intensity modulation depends linearly on the phase change and thus makes the phase object visible.

To improve the contrast, we must decrease the amplitude of the light bypassing the object and make it comparable with the amplitude of light passing through the object. Let us denote the reduced amplitude by $\varepsilon' \ll 1$. The intensity of the diffraction pattern observed in the image plane then becomes

$$2I = |i\varepsilon' + \varepsilon e^{i\delta}|^2 = \varepsilon^2 + \varepsilon'^2 + 2\varepsilon\varepsilon' \sin\delta. \qquad (36.28)$$

This leads to the following expression for the visibility

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{\varepsilon\varepsilon' (\sin\delta_{\max} - \sin\delta_{\min})}{\varepsilon^2 + \varepsilon'^2 + \varepsilon\varepsilon' (\sin\delta_{\max} + \sin\delta_{\min})}, \qquad (36.29)$$

where $\delta_{\max}$ and $\delta_{\min}$ denote the maximum and minimum phase changes in the object. For $\varepsilon' = \varepsilon$, the visibility and contrast of the image considerably increase.

If the phase of the light bypassing the object is changed by $3\pi/2$, i.e. if the complex amplitude of the attenuated light now becomes equal to $\varepsilon' \exp(3\pi/2) = -i\varepsilon'$, we obtain the following relation instead of (36.28):

$$2I = \varepsilon^2 + \varepsilon'^2 - 2\varepsilon\varepsilon' \sin\delta, \qquad (36.30)$$

which gives an equal and opposite contrast, i.e. the dark regions will now appear bright, and vice versa.

In the practical application of the phase-contrast method, the light bypassing the object is attenuated, and its phase changes by $\pi/2$ or $3\pi/2$ not in the object plane but in the focal plane. For this purpose, the central part in the focal plane is covered by a small plate (Fig. 191) which simultaneously decreases the amplitude of light to about ε' and changes its phase by $\pi/2$ or $3\pi/2$. The thickness of the plate is chosen experimentally from the best contrast.

The phase-contrast method was developed in 1953 by Zernike and served as an important step towards holography.

Example 36.1. Find the resolving power of the eye. The radius of the entrance pupil is $R = 1.5$ mm and the refractive index n of the aqueous humor is approximately equal to the refractive index of water.

If the entire space were filled with a liquid having a refractive index equal to that of the aqueous humor, the resolving power of the eye would be given by formula (36.18) in which the wavelength in vacuum is replaced by the wavelength λ' in the liquid:

$$\theta_0' = 0.61\lambda'/R, \tag{36.31}$$

where $\lambda' = \lambda/n$. It should be noted that θ_0' is the angle between the rays at the centre of the bright spot and at its boundary within the eye, and is equal to the minimum angle between the rays entering the eye from a medium filled with the liquid. It is well known from geometrical optics that if media with refractive indices n and n' abut the opposite surfaces of a lens, the acute angle θ between the neighbouring rays changes after the passage of the rays through the lens and becomes

$$\theta' = \theta n/n'. \tag{36.32}$$

If the medium surrounding the eye has $n' = 1$ (air), the angle (36.31) increases n-fold, and hence the minimum resolvable angle θ_0 between the rays incident on the eye also increases n-fold:

$$\theta_0 = n\theta_0' = 0.61\lambda/R. \tag{36.33}$$

Thus, the resolving power of the eye does not depend on the refractive index of the aqueous humor. For $\lambda = 550$ nm, we obtain $\theta_0 = 2.2 \times 10^{-4} = 0.8'$. It is worth noting that if we proceed from the structure of the perceptive elements of the eye (cones and rods), the minimum resolvable angle will again be about $1'$. Consequently, the various elements of the eye are optimally matched.

It should be emphasized that the "economy" in this example indicates just the optimal matching of the various elements of the eye and the absence of additional unexploited possibilities, and by no means the "economy" in creating these possibilities and a reliable functioning of the system as a whole. The reliability of vital systems is always ensured fully, often through a duplication of functions, since without reliability an organism cannot participate in natural selection and develop further.

SEC. 37. SPATIAL FILTERING OF IMAGES

The essence of spatial filtering of images and its applications are discussed.

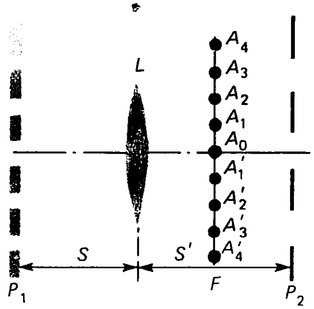


Fig. 192. Explaining the essence of the spatial filtering of an image.

!

The spatial filtering of images is based on the possibility of changing an image by altering the diffraction pattern of the object in the focal plane of the lens in a certain way.

?

Which intensity maxima in the diffraction pattern from the rectangular grating in the Abbe-Porter experiment must be suppressed so that the image is formed only by horizontal fringes? Which intensity maxima in the interference pattern contain information about details of the object?

THE ESSENCE OF SPATIAL FILTERING OF IMAGES. The diffraction formation of an image can be split into two stages (see Sec. 36):

(1) the formation of a diffraction pattern of the object in the focal plane of a lens, and

(2) the transformation of the diffraction pattern in the focal plane of the lens into an image of the object in the image plane.

The entire information contained in the image of the object is also contained in the diffraction pattern of the object in the focal plane of the lens. If the diffraction pattern in the focal plane is altered, for example, if some maxima are eliminated or attenuated, the image of the object will change accordingly. The variation in the image of an object through a modification of its diffraction pattern from which the image is subsequently formed is called the **spatial filtering of the image.**

SPATIAL FILTERING OF THE DIFFRACTION GRATING IMAGES. The essence of spatial filtering can be illustrated graphically by considering the image of a diffraction grating.

A diffraction grating P_1 (Fig. 192), whose distance s from the lens is more than the focal length of the lens, is illuminated by a parallel beam of rays. In plane P_2, whose distance s' from lens L is determined by relation (23.36b), a real inverted image of the grating is produced by the lens in the form of dark and bright fringes with quite sharp boundaries. The sharpness of the boundaries depends on the width of the transparencies and the opacities separating them: the greater the width, the sharper the boundary. In focal plane F of the lens, a Fraunhofer diffraction pattern is formed by the grating. When observed on a real plane surface, this pattern is in the form of diffraction fringes (see Sec. 33). If this surface is imaginary and the rays pass through it without any hindrance, an image of the grating is formed in plane P_2.

The principal intensity maxima of the diffraction pattern are shown in Fig. 192 by $A_0, A_1, A_2, \ldots, A_1', A_2', \ldots$. The most complete image of the grating is obtained in the case when the entire diffraction pattern in plane F participates in the image formation in plane P_2. If a part of the maxima is suppressed, the image deteriorates, some of the details will be missing, and the image of the grating may even be completely distorted. For example, let us eliminate all maxima except A_0. In other words, we reduce the diffraction pattern to just one central maximum. The maximum corresponds to the ray which

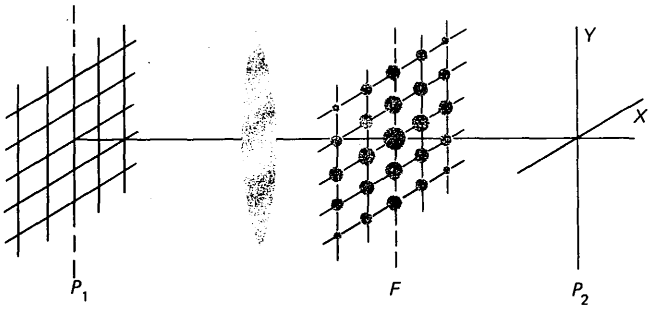


Fig. 193. Schematic diagram of Abbe-Porter experiment.

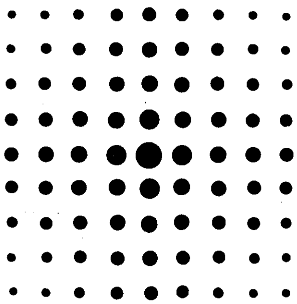


Fig. 194. Diffraction pattern formed by a spatial grating.

encounters no obstacles in its way. A uniform illumination is observed in plane P_2, the image of the grating is invisible, as if it does not exist. Next, let us eliminate all odd maxima A_1, A_3, ..., A_1', A_3', ... The remaining maxima A_0, A_2, A_4, ..., A_2', A_4', ... correspond to a diffraction pattern formed by a grating whose period is half that of the original grating. Hence the image plane has an image of a diffraction grating with a period half that of the original grating, i.e. we get a denser grating. The first-order maxima outline the coarse details of the object, while information about finer details is obtained with the help of the higher-order maxima. If all maxima except the zero- and first-order maxima are suppressed, the distribution of the light field amplitudes in the image plane follows a harmonic law [see (33.58)].

The change in an image produced not by a direct action but by influencing the distribution of amplitudes which ultimately result in the final image forms the basis of spatial filtering. The main problem consists in creating a filter which can appropriately modify the wave passing through it.

ABBE-PORTER EXPERIMENT. The first experiment demonstrating the possibility of spatial filtering of images was performed by Abbe and Porter. The object was a two-dimensional grating formed by mutually perpendicular lines (Fig. 193). The diffraction pattern produced in the focal plane of the lens is shown in Fig. 194. This diffraction pattern forms the image of the grating in the image plane.

If an opaque plate with a central aperture is placed in the focal plane to suppress all maxima except the central one and the maxima which are on the same line with it, the line being parallel to a ruling on the grating (Fig. 195), the image will show only the perpendicular system of lines in the grating. If the central maximum is closed, the image of the grating will have the opposite contrast: the dark lines will be transformed into bright lines, and vice versa.

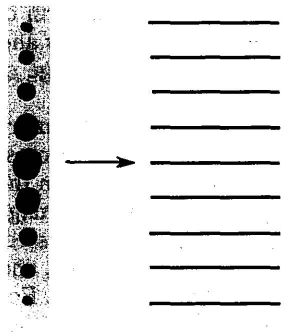


Fig. 195. Filtering of the image of a spatial grating.

SEC. 38. HOLOGRAPHY

The principles of obtaining thin and thick holograms are described as well as the methods of reconstructing the images. Applications of holography are discussed.

SYNCHRONOUS DETECTION. A photographic plate registers the mean square value of the electromagnetic field strength of a light wave, i. e. it records the intensity of the wave. The information concerning the phase of the wave is lost in this process. Thus, a photograph conveys only a limited information about the object. In particular, there is no information about the distances of various parts of the object from the photographic plate, and about several other characteristics. In other words, an ordinary photograph does not allow a complete reconstruction of the wave front registered on it. An ordinary photograph contains a more or less precise information about the amplitudes of the wave registered by the photographic plate, but does not give any information about the phases.

Holography helps in removing this drawback of the ordinary photograph and in recording on a photographic plate information about the amplitudes of the waves incident on it, as well as their phases. In other words, complete information about the waves forming the image can be recorded by using holographic technique. The wave reconstructed with the help of such a recording is completely identical to the original wave and contains complete information carried by the original wave. Hence the method is called holography, i. e. complete recording of a wave.

The idea behind the method is similar to synchronous detection which is frequently used in radio engineering. The information is conveyed in the form of a modulated signal with a high carrier frequency

$$V_m = V_{0m}[1 + mf(t)]\cos\omega t, \tag{38.1}$$

where ω is the carrier frequency, $f(t)$ is the signal to be transmitted, and m is the modulation depth (Fig. 196).

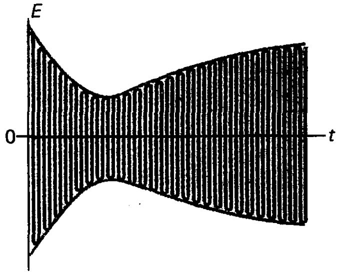


Fig. 196. Amplitude-modulated signal.

The modulated signal (38.1) contains information about the amplitude- and phase spectrum of the function $f(t)$. This information can be separated with the help of synchronous detection.

With the help of certain devices in radio engineering, the electric field strengths of the signals supplied to them can be multiplied and the product obtained at the output. If the modulated signal (38.1) is supplied to such a device together with a constant signal from the carrier frequency

$$V_c = V_{0c}\cos\omega t, \tag{38.2}$$

we shall obtain the following signal at the output

$$V_{out} = \alpha V_m V_c = \alpha V_{0m} V_{0c}[1 + mf(t)](1 + \cos 2\omega t)/2, \tag{38.3}$$

where α is a constant. With the help of a suitable filter, the high-frequency signal is eliminated, and the residual low-frequency signal

$$V^{(c)}_{\text{out}} = A[1 + mf(t)] \tag{38.4}$$

contains complete information about the amplitude and phase spectra of the modulated signal.

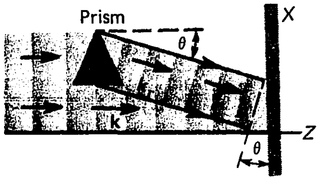


Fig. 197. Diagram of hologram recording of a plane wave.

It is logical to use a method similar to synchronous detection in optical frequency region. For this purpose, we must record the information contained in a light wave as a wave modulated by a certain carrier frequency. When the modulated wave is registered on a photographic plate, the information about the phase of the wave is preserved. **By the method of synchronous detection of the modulated wave recorded on the photographic plate, i.e. by exposing the recorded modulated wave to the carrier frequency, the original wave can be reconstructed. Thus the problem of complete reconstruction of the wave front is solved in principle.**

In order to realize this method in optical region, we must use radiation with a very high degree of coherence. Such a radiation is obtained with the help of lasers. Hence a practical realization of holography became possible only after the creation of lasers capable of producing highly coherent radiation.

The photographic plate containing information about a modulated light wave is called a **hologram. The aim of holography is to work out methods for recording holograms and for reconstructing the wave front from holograms.**

The idea of holography was first put forth in 1920 by the Polish physicist Wolfke (1883-1947), but was soon forgotten. The idea was independently put forward and developed by the British physicist D. Gabor who was awarded the Nobel Prize in Physics in 1971.

HOLOGRAM OF A PLANE WAVE. The wave with the carrier frequency is called the reference wave, while the wave containing information about the object is called the signal wave.

The plane signal wave of frequency ω (Fig. 197) propagates in the positive direction of the Z-axis. In the XY-plane perpendicular to the Z-axis, we place a photographic plate. The reference wave is formed by dividing the wave front, and is directed with the help of a prism P towards the photographic plate where it overlaps the signal wave which is also obtained by dividing the wave front. The angle of inclination of the reference wave to the Z-axis is denoted by θ.

The signal and reference waves are written in the form

$$E_s = E_1 e^{-i(\omega t - kz)}, \tag{38.5}$$

$$E_r = E_0 e^{-i(\omega t - k_z z - k_x x)}. \tag{38.6}$$

It should be noted that relations (38.5) and (38.6) are correct but for the constant phase. It could be assumed that E_1 and E_0 in these formulas are complex and contain these phases albeit not in an explicit form. However, this only complicates the form of notation in the subsequent formulas and does not contain any significant information. Hence we shall assume that E_1 and E_0 are real.

Considering that $k_x = k \sin\theta$ and $k_z = k\cos\theta \approx k$ up to second-order terms in θ which are assumed to be small, we can write (38.6) in the form

$$E_r = E_0 e^{-i(\omega t - kz - kx\sin\theta)}. \tag{38.7}$$

The total amplitude of the electric field strength in the plane of the photographic plate is given by

$$E = E_s + E_r = e^{-i(\omega t - kz)}(E_1 + E_0 e^{ikx\sin\theta}), \tag{38.8}$$

where E_1 and E_0 can be assumed to be real amplitudes. This leads to the following expression for intensity distribution:

$$I(x) = (1/2)|E|^2 = (1/2)[E_1^2 + E_0^2 + 2E_0E_1\cos(kx\sin\theta)]. \tag{38.9}$$

According to this relation, the photographic plate in this case also records the phase difference between reference and signal waves, i.e. the phase of the signal wave if we assume that the phase of the reference wave is given.

RECONSTRUCTION OF THE IMAGE. A photographic emulsion consists of silver halide particles suspended in a gelatine base. This emulsion is applied to a glass substrate or an acetate film. When light falls on a silver halide particle, it creates centres of reduced silver, or the development centres. In the developing process, particles containing development centres are reduced to metallic silver. If there are no development centres in a particle, it remains in the form of a halide. During the process of a "fixation" following development, silver halide particles are removed and the plate contains metallic silver only in small particles which form dark spots on the plate. It is shown in the theory of the photographic process that the density of plate blackening is

$$Q = \gamma \log |E|^2, \tag{38.10}$$

where γ characterises the photographic material. After the development, the transmission factor of the plate has the form

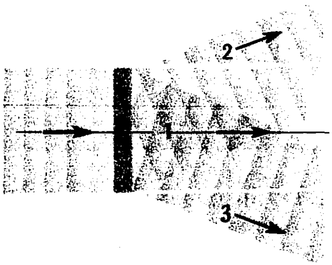


Fig. 198. Reconstruction of a wave recorded on a hologram.

$$\underline{\overline{\tau} = (\overline{|E|^2})^{-\gamma/2}.} \tag{38.11}$$

Combining (38.9) and (38.11), we find that

$$\tau = [E_0^2 + E_1^2 + 2E_0E_1\cos(kx\sin\theta)]^{-\gamma/2}. \tag{38.12}$$

Since $E_1 \ll E_0$ under normal conditions, instead of (38.12) we can write

$$\begin{aligned}\tau &= E_0^{-\gamma}[1 - \gamma E_1^2/2E_0^2 - (\gamma E_1/E_0)\cos(kx\sin\theta)]\\ &= (E_0^{-\gamma-2}/2)[2E_0^2 - \gamma E_1^2 - 2\gamma E_0E_1\cos(kx\sin\theta)].\end{aligned} \tag{38.13}$$

Neglecting the scale factor $(E_0^{-\gamma-2}/2)$ which is of no significance in the subsequent analysis, we can write (38.13) in a more convenient form:

$$\tau = 2E_0^2 - \gamma E_1^2 - \gamma E_0E_1 e^{ikx\sin\theta} - \gamma E_0E_1 e^{-ikx\sin\theta}. \tag{38.14}$$

We direct towards the hologram a plane wave along the path of the signal wave (Fig. 198):

$$E_{\text{rec}} = E_2 e^{-i(\omega t - kz)}. \tag{38.15}$$

The light field emerging from the hologram and given by

$$\begin{aligned}E_{\text{out}} = E_{\text{rec}}\tau = E_2(2E_0^2 - \gamma E_1^2)e^{-i(\omega t - kz)}\\ - \gamma E_0E_1E_2 e^{-i(\omega t - kz + kx\sin\theta)} - \gamma E_0E_1E_2 e^{-i(\omega t - kz - kx\sin\theta)},\end{aligned} \tag{38.16}$$

consists of three plane waves (Fig. 198):

(1) the first term in (38.16) describes a plane wave *1* propagating in the positive direction of the *Z*-axis as a continuation of the wave incident on the hologram;

(2) the second term describes a wave *2* propagating at an angle θ to the *Z*-axis inclined towards the positive direction of the *X*-axis. This can be seen from a comparison of the signs of $kx\sin\theta$ in the exponents of wave *2* and the wave described by Eq. (38.7) and propagating towards the negative direction of the *X*-axis; and

(3) the third term describes a wave *3* propagating at an angle θ to the *Z*-axis and inclined towards the negative direction of the *X*-axis.

The light field represents the diffraction of a plane wave incident on the hologram. It can be seen that only the first-order diffraction is observed, as should indeed be the case when the transmission coefficient (38.14) varies according to a harmonic law [cf. (33.64)].

HOLOGRAM OF A POINT OBJECT. The method of obtaining such a hologram is shown in Fig. 199. The reference wave is formed after refraction in the prism. This wave is defined by the following formula which is analogous to (38.7):

$$E_r = E_0 e^{-i(\omega t - kz - kx\sin\theta)}. \tag{38.17}$$

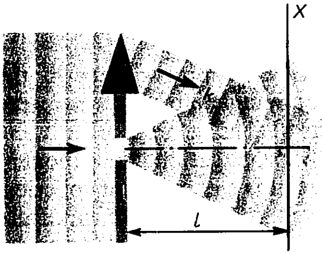


Fig. 199. Obtaining the hologram of a point object.

The spherical wave emanating from the point source can be presented in the plane of the photographic plate:

$$E_{\mathrm{sph}} = E_1\,\mathrm{e}^{-i[\omega t - kz - kx^2/(2z_0)]}, \tag{38.18}$$

the slight variation in the amplitude of the electric field strength with increasing distance from the point source can be neglected as we move along the photographic plate.

It should be observed that the constant phases are not written in explicit form in formulas (38.17) and (38.18) since they have no significant physical meaning. The existence of a constant phase difference between waves leads only to a small displacement of the diffraction pattern in space, without causing any noticeable change in the pattern itself. Hence there is no need to take into account this constant phase difference. Therefore, the origin of the coordinate system is not important in the expression for plane waves. However, for the spherical wave (38.18) presented through the exponential term $ikx^2/(2z_0)$, it must be remembered that the origin of coordinates coincides with the centre of the aperture in the screen. Hence the denominator of this expression contains z_0 to emphasize that we are interested in the diffraction pattern in the plane of the photographic plate at a distance z_0 from the aperture.

The total electric field strength on the plate

$$E = E_{\mathrm{r}} + E_{\mathrm{sph}} \tag{38.19}$$

assumes the form

$$E = \mathrm{e}^{-i(\omega t - kz)}\,[E_0\mathrm{e}^{ikx\sin\theta} + E_1\mathrm{e}^{ikx^2/(2z_0)}], \tag{38.20}$$

whence

$$\begin{aligned}|E|^2 &= E_0^2 + E_1^2 + E_0E_1\,\mathrm{e}^{i[kx\sin\theta - kx^2(2z_0)]} + \mathrm{e}^{-i[kx\sin\theta - kx^2/(2z_0)]}\\ &= E_0^2 + E_1^2 + 2E_0E_1\cos[kx\sin\theta - kx^2/(2z_0)].\end{aligned} \tag{38.21}$$

It can be seen that the phase relations between the waves are registered on the plate. As in (38.19), we obtain the following expression for the transmission coefficient:

$$\begin{aligned}\tau &= 2E_0^2 - \gamma E_1^2 - \gamma E_0E_1\,\mathrm{e}^{i[kx\sin\theta - kx^2/(2z_0)]}\\ &\quad - \gamma E_0E_1\,\mathrm{e}^{-i[kx\sin\theta - kx^2/(2z_0)]}.\end{aligned} \tag{38.22}$$

Hence when the hologram is illuminated by a plane wave

$$E_{\mathrm{rec}} = E_2\,\mathrm{e}^{-i(\omega t - kz)} \tag{38.23}$$

the light field at the outlet of the hologram is given by an expression analogous to (38.16):

$$\begin{aligned}E_{\mathrm{out}} &= E_{\mathrm{rec}}\tau = E_2(2E_0^2 - \gamma E_1^2)\,\mathrm{e}^{-i(\omega t - kz)}\\ &\quad - \gamma E_0E_1E_2\,\mathrm{e}^{-i[\omega t - kz - kx\sin\theta + kx^2/(2z_0)]}\\ &\quad - \gamma E_0E_1E_2\,\mathrm{e}^{-i[\omega t - kz + kx\sin\theta - kx^2/2z_0)]}.\end{aligned} \tag{38.24}$$

This light field consists of three waves (Fig. 200):

(1) the first term represents a plane wave propagating in the same direction as the wave incident on the hologram;

(2) the second term presents a wave propagating in the positive direction of the X-axis at an angle θ to the Z-axis (the exponent contains the term $ikx \sin\theta$). This is a spherical wave which is concave in the direction of propagation [the exponent contains the term $-ikx^2/(2z_0)$], i. e. it is a convergent spherical wave converging at the focus located at the centre of curvature of the surface. The focus serves as the real image of the point object. It can be seen that it is the mirror image of the point object recorded on the hologram if the plane of the mirror coincides with the plane of the hologram, and the point object is situated relative to the hologram in the same way as during the recording of the hologram (Fig. 201);

(3) the third term in (38.24) represents a wave propagating in the positive direction of the X-axis at an angle θ to the Z-axis (the exponent contains the term $-ikx \sin\theta$). This wave is spherical with its concave face against the direction of propagation [the exponent contains the term $ikx^2/(2z_0)$], i. e. it is a divergent spherical wave. It is absolutely identical to the spherical wave propagating from the point object during the recording of the hologram, and produces a virtual image of the point at the position where the point object was located during the hologram recording. If the wave falls on the human eye, it produces the same perception as produced by a spherical wave diverging from the point object itself. Thus, the holographic recording allows a complete construction of the wave front.

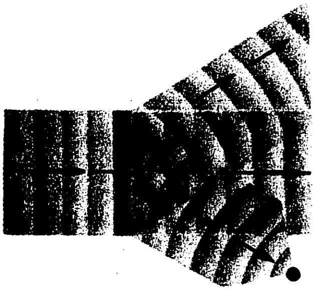


Fig. 200. Reconstruction of the wave front from a point object.

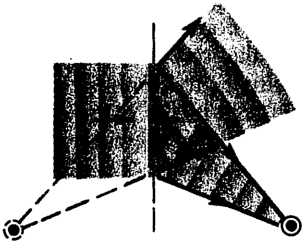


Fig. 201. Real and virtual images of a point formed during reconstruction of a hologram.

HOLOGRAM OF AN ARBITRARY OBJECT. The wave front of an arbitrary object is composed of the wave fronts formed by its points (see Secs. 31, 34, 36). Hence the recording of the hologram of an arbitrary object is carried out in the same way as for a point object, except that the signal wave carrying information about the object must be coherent with the reference wave. For this purpose, the object is illuminated by a wave coherent with the reference wave. This can be done with the help of various methods.

Figure 202 shows one of the possible methods of recording a hologram. It can be seen that the reference wave is formed by the mirror reflection of the wave used for illuminating the object. The wave arriving from the object to the photographic plate can be presented in the plane of the plate in the following form:

$$E_{\text{sph}}(x, y) = E_1(x, y)\, e^{-i[\omega t - kz - \varphi(x, y)]}, \tag{38.25}$$

where $E_1(x, y)$ and $\varphi(x, y)$ are the amplitude- and phase distributions of the light field from the source. The reference wave is presented in the form (38.19). The square of the

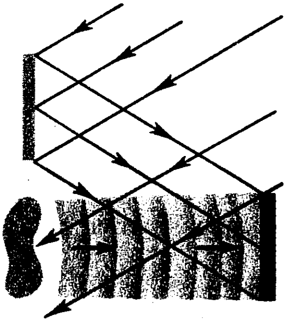


Fig. 202. A set-up for recording holograms.

absolute value of the electric field strength recorded on the hologram is given by the following expression instead of (38.21):

$$|E|^2 = E_0^2 + E_1^2 + E_0E_1\left(\mathrm{e}^{i(kx\sin\theta-\varphi)} + \mathrm{e}^{-i(kx\sin\theta-\varphi)}\right). \tag{38.26}$$

When the hologram is illuminated by a monochromatic wave, we obtain an expression similar to (38.24) for the reconstructed wave:

$$E_{\text{out}} = E_{\text{rec}}\tau = E_2(2E_0^2 - \gamma E_1^2)\,\mathrm{e}^{-i(\omega t-kz)} - \gamma E_0E_2E_1\,\mathrm{e}^{-i(\omega t-kz-kx\sin\theta+\varphi)} - \gamma E_0E_2E_1\,\mathrm{e}^{-i(\omega t-kz+kx\sin\theta-\varphi)}. \tag{38.27}$$

Like (38.24), this relation describes the transmitted wave and the waves producing the real and virtual images of the object (Fig. 203).

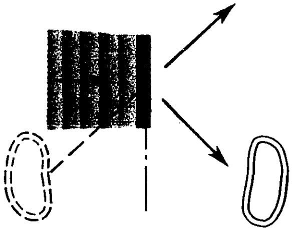


Fig. 203. Reconstruction of real and virtual images of an object.

The wave producing the virtual image of the object is an exact reconstruction of the wave emanating directly from the object. This virtual image is three-dimensional and hence we can have a lateral view of the object by changing the angle of view. The sides of the object can be seen by laterally moving the head. The object can also be photographed from different aspects under the condition, of course, that the objective of the photographic camera remains within the reconstructed wave.

During the recording of the hologram, the signal wave and the reference wave must be mutually coherent. In any case, the coherence width must not be less than the size of the object, and the coherence length must not be less than the path difference between the signal wave and the reference wave. Under real conditions, this means that the radiation used in hologram recording must have a high degree of spatial and time coherence. This requirement is met by laser radiation. The reconstruction of holograms is also carried out with the help of lasers. However, the frequency of the laser radiation used for reconstruction of the hologram may differ from the frequency used while recording the hologram. This is due to the fact that the reconstruction of a hologram boils down to the diffraction of the wave incident on the hologram. The diffraction angle increases with wavelength. Hence when a hologram is reconstructed by using radiation of wavelength larger than that used for the recording, the image becomes larger than the object.

The reconstruction of a hologram can also be carried out without lasers. A very small incoherent object, seen under the coherence angle (27.24), produces radiation with a fairly high degree of coherence at the coherence width (27.25). For example, at a distance of a stretched arm, a luminous pinhead creates a highly coherent light field on the pupil of the eye.

Consequently, if a hologram is placed between a luminous pinhead and the eye (the pinhead is at an arm's length from the eye), the holographic image of the object recorded on the hologram can be reconstructed. The image reconstructed in this way is not so sharp as the one obtained with the help of laser radiation, i. e. higher-order diffraction maxima are lost in this case. The three-dimensional nature of the image, however, remains intact.

QUALITY OF A PHOTOGRAPHIC PLATE AND THE EXPOSURE TIME. As in the case of conventional photography, a photographic plate used for recording a hologram registers the intensity of the luminous flux. In other words, it performs the same function in both cases. The only difference is that the hologram registers much finer details of the intensity distribution and a much wider range of intensity variation than a conventional photograph.

The photographic plate must ensure a recording of the diffraction pattern forming the hologram. The condition for maxima in the hologram of a plate wave (38.9) has the form

$$\cos(kx\sin\theta) = 1, \tag{38.28}$$

while the separation Δx between the maxima is defined by the relation

$$k\Delta x\sin\theta = 2\pi. \tag{38.29}$$

Hence it follows that $\Delta x = \lambda/\sin\theta$. For example, if $\theta = 15°$, $\sin\theta = 0.26$ and hence $\Delta x \simeq 4\lambda \simeq 2$ µm, i. e. the photographic plate must be capable of resolving lines that are 2 µm apart. The resolving power of photographic plates is usually expressed in terms of lines per millimetre that can be resolved by it. In the present case, the resolving power of the photographic plate must be 500 lines/mm. It is desirable to have plates with a still higher resolving power. For this purpose, we must use very fine grains of silver halide, which reduces the sensitivity of the plate. Hence plates with a high resolving power have a low sensitivity and require a longer exposure time, running into several seconds for lasers of moderate power. During exposure, the process must be stationary and there should be no relative movement between the object and the instrument to within a fraction of a wavelength (usually $\lambda/4$). When high-power pulsed lasers are used, the exposure time can be reduced to the pulse duration (milliseconds and less). Holograms of rapidly moving objects can be recorded under these conditions.

THREE-DIMENSIONAL REPRODUCTION OF THE OBJECT. The virtual image obtained with the help of a hologram is seen as its three-dimensional photograph. In a certain sense, the real image of the object is its three-dimensional reproduction but its parallax relations are opposite to those of the real object. The object appears to be hanging in air and can be photographed. Thus, a spatial reconstruction of the situation is possible, i. e. an illusion can be created that there is a certain object in a room although it is actually not there.

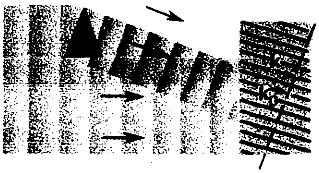


Fig. 204. Hologram recording of a plane wave in a thick emulsion layer.

THICK HOLOGRAMS (DENISYUK'S METHOD). In contrast to the method of recording holograms on ordinary photographic plates with a thin layer of emulsion, the Soviet scientist Yu. N. Denisyuk (born 1927) proposed in 1962 the method of thick holograms in which the interference pattern of the diffracted rays is not two-dimensional but three-dimensional and covers the entire thickness of the emulsion.

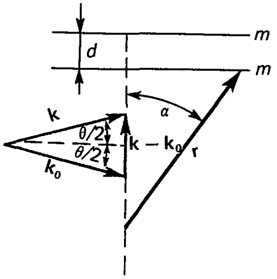


Fig. 205. To the determination of surfaces with maximum darkening in a thick emulsion layer.

Let us analyze a thick hologram of a plane wave which can be recorded by using the set-up shown in Fig. 197 in which a conventional photographic plate is replaced by a thick emulsion layer (Fig. 204). We denote by $\mathbf{k}$ and $\mathbf{k}_0$ the wave vectors of the signal and reference waves. Like (38.9), the square of the amplitude of the field strength can be written as

$$|E|^2 = |E_0 e^{i\mathbf{k}_0\mathbf{r}} + E_1 e^{i\mathbf{k}\mathbf{r}}|^2$$
$$= E_0^2 + E_1^2 + 2E_0E_1 \cos[(\mathbf{k}_0 - \mathbf{k})\cdot\mathbf{r}]. \tag{38.30}$$

The condition for the maximum values of $|E|^2$ can be written in the form

$$(\mathbf{k}_0 - \mathbf{k})\cdot\mathbf{r} = 2\pi m \quad (m = 0, \pm 1, \pm 2, \ldots). \tag{38.31}$$

This equation represents a system of planes perpendicular to the vector $\mathbf{k}_0 - \mathbf{k}$ (Fig. 205). The distance d between the planes satisfies the following condition based on (38.31):

$$|\mathbf{k}_0 - \mathbf{k}|\, d = 2\pi. \tag{38.32}$$

Considering that $|\mathbf{k}| = |\mathbf{k}_0|$, we obtain

$$|\mathbf{k} - \mathbf{k}_0| = 2k \sin(\theta/2). \tag{38.33}$$

This gives [see (38.32)]

$$d = \lambda/[2 \sin(\theta/2)]. \tag{38.34}$$

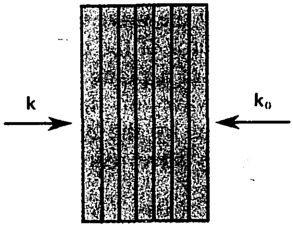


Fig. 206. Spatial interference pattern for the case of two waves propagating from opposite directions.

In particular, when two waves coming from opposite directions ($\theta = \pi$) interfere, the planes of maximum darkening are parallel to the wave fronts of the interfering waves (Fig. 206). It can be concluded on the basis of (38.34) that the distance between the planes is $d = \lambda/2$.

BRAGG'S LAW. Waves are partially reflected and partially transmitted through each plane of maximum darkening in which the maximum particle-density of reduced silver is concentrated. However, reflection is possible from a system of parallel planes only if the waves reflected from adjacent planes augment each other (see Sec. 29). The angle of incidence α at which the reflection from a system of parallel planes takes place (Fig. 207) is defined in analogy with (29.4) and (29.5) through the condition

$$2d\cos\alpha = m\lambda, \tag{38.35}$$

where the angle of refraction θ_{refr} is equal to the angle of incidence, and the refractive index of the medium between the reflecting surfaces is equal to unity. Equality (38.35) is called Bragg's law. It was formulated by W. L. Bragg (1890-1971) and independently by the Russian crystallographer Yu. V. Wulf (1863-1925).

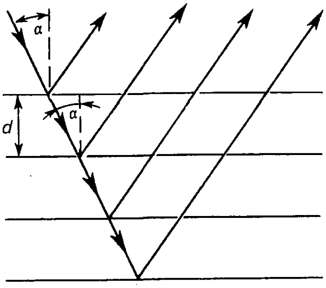


Fig. 207. To the derivation of Bragg's law.

RECORDING OF HOLOGRAMS AND RECONSTRUCTION OF A PLANE WAVE. Figure 208 shows the hologram of a plane wave whose wave vector **k** forms an angle $\pi - \beta$ with the wave vector $\mathbf{k}_0$ of the reference wave. The planes of maximum darkening are situated perpendicular to the direction of the vector $\mathbf{k}_0 - \mathbf{k}$. The normal to the surfaces forms an angle 2β with the vector **k**, and the distance between the surfaces is given, in accordance with (28.34), by the relation

$$d = \lambda/\{2\sin[(\pi - 2\beta)/2]\} = \lambda/(2\cos\beta).$$

This gives [see (38.35)]

$$[2\lambda/(2\cos\beta)]\cos\alpha = m\lambda \tag{38.36}$$

or

$$\cos\alpha = m\cos\beta. \tag{38.37}$$

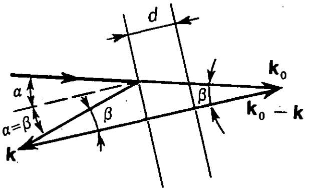


Fig. 208. Reflection of a plane wave at a thick hologram.

It follows from here that only the first-order reflection takes place ($m = 1$) and $\alpha = \beta$. Consequently, only the wave with the wave vector $\mathbf{k}_0$ is reflected, and the wave vector of the reflected wave is **k**. In other words, if the hologram of a plane wave is recorded with the help of a reference wave of the same frequency, the illumination of the hologram by the reference wave leads to a reconstruction of the plane wave recorded in the hologram.

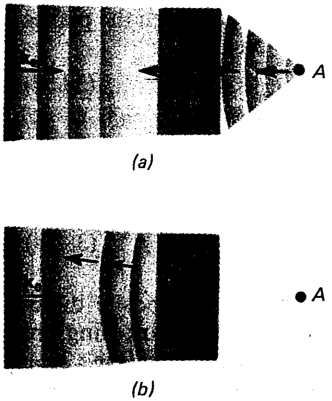


Fig. 209. Recording of a thick hologram of a point object (*a*) and reconstruction of a wave (*b*).

RECORDING OF HOLOGRAMS AND RECONSTRUCTION OF A SPHERICAL WAVE. In a small region away from the source, a spherical wave can be considered as a plane wave. Hence, by exposing a thick photographic plate and a point object *A* to the same reference wave with wave vector $\mathbf{k}_0$ (Fig. 209), we obtain in the bulk of the emulsion a set of surfaces of maximum darkening. The distance between the surfaces satisfies condition (38.34) with $\theta = \pi - 2\beta$. It follows from what has been said about Eq. (38.37) that the exposure of a hologram to a plane wave with a wave vector $\mathbf{k}_0$

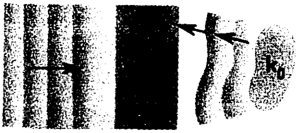


Fig. 210. Recording of a thick hologram of an arbitrary object.

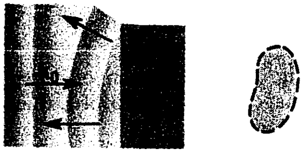


Fig. 211. Reconstruction of the image of an arbitrary object.

leads to a complete reconstruction of the spherical wave recorded on the hologram as a result of the reflection of a plane wave from the diffraction pattern formed in the bulk of the emulsion during the hologram recording.

HOLOGRAPHIC RECORDING AND RECONSTRUCTION OF THE IMAGE OF AN ARBITRARY OBJECT. When a photographic plate and an object are exposed to the same wave (Fig. 210), each point on the object forms a diffraction pattern in the bulk of the emulsion as described above (Fig. 209). The aggregate of the diffraction patterns from all the points on the object forms the hologram.

The reconstruction of the image is carried out by exposing the hologram to a wave coincident with the reference wave used during the hologram recording (Fig. 211). The image of the object is virtual and is in the position where the real object was during the hologram recording. Such a reconstruction of the image has a significant drawback: the reconstructing wave coincides in space with the reconstructed wave. In order to eliminate this drawback, we can illuminate the hologram by a wave with a wave vector $\mathbf{k}_0$ directed at an appropriate angle to the surface of the hologram (Fig. 212). The wave scattered from the diffraction pattern is observed at the angle of reflection which is equal to the angle of incidence. Consequently, the reconstructing and the reconstructed waves become separated in space. The image of the object is virtual and its position depends on the angle at which the hologram is illuminated.

Since the distance d between the surfaces of maximum darkening is roughly equal to $\lambda/2$, we conclude that $m = 1$ in (38.35). There is no need to illuminate the hologram by monochromatic light for reconstructing the image. When the hologram is illuminated by white light, only the wave satisfying condition (38.35) is reflected. This simplifies the reconstruction of thick holograms and also makes it possible to obtain coloured images.

A thick hologram can be used to produce a real image of an object. For this purpose, we must illuminate the hologram from the same side on which the object was situated (Fig. 213) during recording in accordance with the scheme shown in Fig. 210. It can be seen from Fig. 213 that the reconstructed wave is convergent, and hence produces a real image of the object.

THREE-DIMENSIONAL COLOURED IMAGE. It is well known that any colour can be synthesized from three independent colours, which are usually red, green and blue. A coloured image is formed by combining images in three colours.

In order to obtain a coloured image, the object is exposed

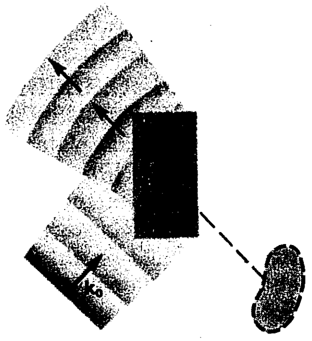


Fig. 212. Separation of the reconstructing and the reconstructed waves.

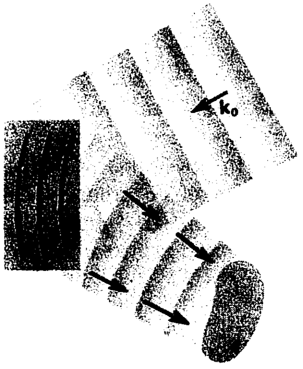


Fig. 213. Formation of a real image with the help of a thick hologram.

during recording to waves corresponding to the three primary colours. Three diffraction patterns from the three different waves are recorded on the hologram. The image is reconstructed by using white light. In accordance with (38.35), only the waves having frequencies that were used for recording are reflected, and their intensity matches the intensity of waves during recording. This means that the wave reconstructed with the help of white light produces a three-dimensional coloured image of the object.

The quality of an image depends on the resolving power of the photographic emulsion, on the nature of distortions in the diffraction pattern due to mechanical deformation in the emulsion during development and drying, as well as on other factors. At present, it is possible to produce high-quality holograms.

PECULIARITIES OF HOLOGRAMS AS CARRIERS OF INFORMATION. Radiation from all points of the object is incident on any part of a hologram. This means that a part of the hologram contains all information about the object carried in the entire hologram. If a hologram is divided into several parts, each part allows a reconstruction of the entire image of the object. In other words, a part of the hologram contains the same information as the entire hologram. Hence a deterioration of the quality of recording at some parts of the hologram (for example, rubbed out or ground parts) does not result in a deterioration of the quality of the image. The information recorded on a hologram has a high reliability margin.

A hologram carries incomparably more information than

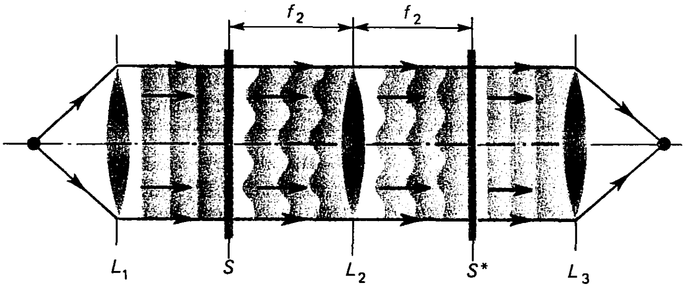


Fig. 214. Spatial filtering of the image for solving the problem of identification of images.

a conventional photograph. The capacity and reliability of a hologram make it a very promising tool for carrying information.

APPLICATIONS OF HOLOGRAPHY. The main task of holography is to produce three-dimensional images. The development of thick-plate holography technique made it possible to create coloured three-dimensional photographs. This idea is being explored further with a view to produce holographic movies, television, etc.

The most significant advances in technological aspects have been made in the field of holographic interferometry. The wave reconstructed from a hologram produces a copy of the object at the instant when the hologram was recorded. If this wave is compared with a wave from the object reconstructed from a hologram recorded at a different instant of time, we can draw conclusions about the change in the object in the period between the two recordings. Since a hologram registers an object with a very high precision, this method allows us to analyze accurately the phenomena which affect a hologram, such as deformation, vibrations, etc. This method is called holographic interferometry. The hologram is influenced not only by a spatial displacement of parts of the object or by its displacement as a whole, but also by the conditions of reflection and refraction of light by the object as well as some other factors which cause an amplitude- or phase modulation of light. Hence holographic interferometry is also used for studying the stress distribution in a body, torques, temperature distribution, etc. Holography can be also used for ensuring the precision of machining of parts.

It is often required to obtain the three-dimensional image of an object which does not exist as yet, and hence optical methods are not in a position to provide a hologram of such an object. In this case, the hologram is obtained with the help of a computer (digital hologram), and the results of computation are then transferred to a photographic plate. The three-

!

The physical meaning of holography is the realization of the principle of synchronous detection, which was known in radiophysics long ago, in the visible range of electromagnetic waves. For this purpose, highly coherent light must be used.

dimensional image of the object from such a computer hologram is then reconstructed by usual optical methods. The surface of the object obtained through a computer hologram is used as the standard, and the surface of the real object being prepared by the appropriate instruments is compared with this standard by holographic interference method. Holographic interferometry allows a comparison of the surface of the finished product with the standard with an extremely high accuracy (up to a fraction of a wavelength). This allows the fabrication of highly complex surfaces with the same precision, which would be impossible without digital holography and holographic interferometry. Naturally, the computer hologram need not be reconstructed by optical means for a comparison of the standard surface with the surface being prepared. It is sufficient to obtain the hologram of the object, transform it into the digital language of the computer and compare with the digital hologram of the standard. In principle, both these methods are identical.

The peculiarities of holograms as carriers of information make it worthwhile to develop the techniques of holographic memory, characterized by a large capacity, reliability, high computation speed, etc.

Holograms can also be used effectively in solving problems related to identification of images. In the physical sense, this problem is reduced to the spatial filtering of images. The plane wave emerging from lens L_1 in Fig. 214 is incident on a diffraction structure S. The signal leaving S has to be identified. A plate with recorded signal is placed at the first focus of lens L_2, at the second focus of which mask S^* for signal S is placed. A mask is a diffraction structure which makes the wave plane upon passing through it. The relation between the recording on S and mask S^* is determined from the fact that the lens transfers the diffraction pattern from one focal plane to the other by means of Fourier transform (see Sec. 35). Hence mask S^* must eliminate the Fourier transform from S (that is why it is denoted by S^*). A detector of light energy is placed at the focus of lens L_3. If S is a signal for mask S^*, the entire energy passing through the lens is concentrated at the focus. If, however, S is not a signal for S^*, the wave emerging from S^* will not be plane, and hence the light energy passing through lens L_3 will not converge at the focus and will be distributed over the focal plane. As a result, the concentration of the energy arriving at the focus will not be detected. It can therefore be concluded that S does not contain the signal that is being sought with the help of mask S^*. Thus, the identification procedure involves the "presentation" of different images

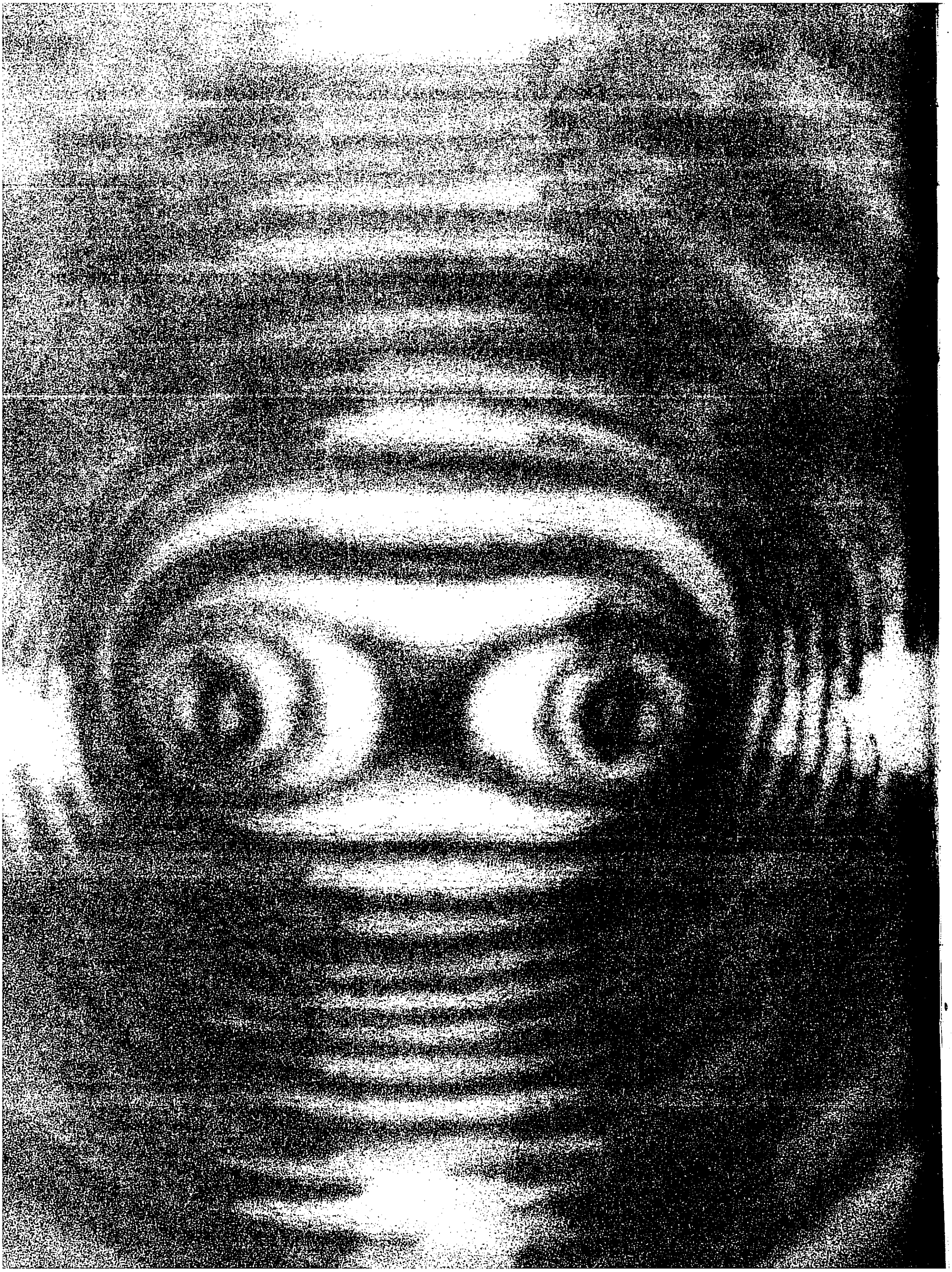
?

Can thin holograms be reconstructed with the help of conventional, incoherent light sources? How can this be done? Which changes take place in the image if a hologram is reconstructed with the help of a radiation with a wavelength different from that in which it was recorded?
What physical factor makes it possible to reconstruct an image recorded on a thick hologram with the help of a radiation having a continuous spectrum?

for S and selection of the one corresponding to the image recorded in mask S^*.

This simplified diagram can be improved. For example, if the signal is supplied to all the masks to which it may correspond, the mask matching the signal will “respond” to it. Thus, the image recorded in S will be “identified”.

In the schematic diagram (Fig. 214) for spatial filtration for identification of images, S and S^* may be holograms. This enhances the possibilities of the method since the information capacity of holograms is much larger than that of the photographs.



Chapter **8**

Propagation of Light in Anisotropic Media

Basic idea:
The electric properties of a medium are characterized by the permittivity tensor. A transition to the principal axes of the tensor makes the analysis of light propagation in all anisotropic medium visual.

SEC. 39. ANISOTROPIC MEDIA

The anisotropy of a medium is defined with the help of the permittivity tensor. A transition is made to the principal tensor axes.

SOURCES OF ANISOTROPY. Optical anisotropy is the dependence of optical properties of a medium on direction. This anisotropy is due to the dependence of dielectric or magnetic properties of the medium on direction. Total anisotropy is the result of the anisotropy in properties of individual atoms and the anisotropy of their ordering in space.

ANISOTROPIC DIELECTRIC MEDIA. The anisotropy of dielectric properties of a medium means that the dependence of the polarization of a medium on the electric field strength cannot be characterized just by one scalar quantity, viz. permittivity. The relations connecting the projections of polarization with the projections of the electric field strength in an anisotropic medium are more complicated than for an isotropic medium:

$$\begin{aligned} P_x &= \varepsilon_0\varkappa_{xx}E_x + \varepsilon_0\varkappa_{xy}E_y + \varepsilon_0\varkappa_{xz}E_z, \\ P_y &= \varepsilon_0\varkappa_{yx}E_x + \varepsilon_0\varkappa_{yy}E_y + \varepsilon_0\varkappa_{yz}E_z, \\ P_z &= \varepsilon_0\varkappa_{zx}E_x + \varepsilon_0\varkappa_{zy}E_y + \varepsilon_0\varkappa_{zz}E_z, \end{aligned} \tag{39.1}$$

where P_x, P_y, P_z are the projections of polarization and ε_0 is the dielectric constant.

The set of quantities $\varkappa_{xx}$, $\varkappa_{xy}$, ... is called the permittivity tensor. For convenience of notation, we shall denote the X-, Y-, and Z-axes by subscripts 1, 2, and 3 respectively. Formula

(39.1) can then be written in the form

$$P_i = \varepsilon_0 \sum_j \varkappa_{ij} E_j. \tag{39.2}$$

The relation between the electric displacement vector D and the polarization P has the form

$$\mathrm{D} = \varepsilon_0 \mathrm{E} + \mathrm{P}. \tag{39.3}$$

This relation is valid for isotropic and anisotropic media. Taking (39.2) into account, we can write this relation in the following form for an anisotropic medium:

$$D_i = \varepsilon_0 E_i + \varepsilon_0 \sum_j \varkappa_{ij} E_j = \sum_j \varepsilon_0 (\delta_{ij} + \varkappa_{ij}) E_j, \tag{39.4}$$

where δ_{ij} is the Kronecker delta. It is convenient to express this relation in the form

$$D_i = \sum_j \varepsilon_{ij} E_j, \tag{39.5}$$

where

$$\varepsilon_{ij} = \varepsilon_0 (\delta_{ij} + \varkappa_{ij}) \tag{39.6}$$

is the permittivity tensor.

The density of electric energy in a medium is given by the formula

$$w = \frac{1}{2} \mathrm{E} \cdot \mathrm{D} = \frac{1}{2} \sum_i E_i D_i = \frac{1}{2} \sum_{i,j} E_i \varepsilon_{ij} E_j, \tag{39.7}$$

where expression (39.5) has been used for D_i. Interchanging the summation indices ($i \leftrightarrow j$), we can write (39.7) in the form

$$w = \frac{1}{2} \sum_{i,j} E_j \varepsilon_{ji} E_i. \tag{39.8}$$

Subtracting (39.8) from (39.7) termwise, we obtain

$$0 = \frac{1}{2} \sum_{i,j} (\varepsilon_{ij} - \varepsilon_{ji}) E_i E_j. \tag{39.9}$$

Since the projections E_i are independent, we conclude that

$$\varepsilon_{ij} - \varepsilon_{ji} = 0, \tag{39.10}$$

!

The electric properties of an anisotropic medium are described by the symmetric permittivity tensor.

i.e. the permittivity tensor is symmetric:

$$\varepsilon_{ij} = \varepsilon_{ji}. \tag{39.11}$$

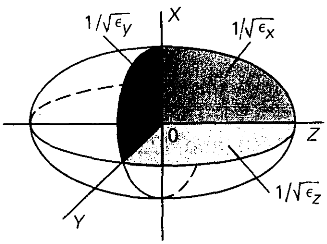


Fig. 215. Principal axes relative to which the permittivity tensor is diagonal.

PERMITTIVITY TENSOR. Since the electric energy density w is positive, the quadratic form on the right-hand side of (39.8) is a positive definite quantity. For further analysis, it is more convenient to use the variables

$$x_i = E_i/\sqrt{2w}, \tag{39.12}$$

and to replace (39.8) by the equality

$$\sum_{i,j} \varepsilon_{ij} x_i x_j = 1. \tag{39.13}$$

It is well known from elementary mathematics that the quadratic form (39.13) can be reduced to

$$\varepsilon_x x^2 + \varepsilon_y y^2 + \varepsilon_z z^2 = 1 \tag{39.14}$$

by an appropriate orientation of the coordinate axes. In the system of coordinates where the quadratic form (39.13) assumes the form (39.14), the permittivity tensor is diagonal:

$$\varepsilon_{ij} = \begin{pmatrix} \varepsilon_x & & 0 \\ & \varepsilon_y & \\ 0 & & \varepsilon_z \end{pmatrix}. \tag{39.15}$$

Equation (39.14) describes an ellipsoid with principal axes $1/\sqrt{\varepsilon_x}$, $1/\sqrt{\varepsilon_y}$, and $1/\sqrt{\varepsilon_z}$ (Fig. 215). The X-, Y- and Z-axes are called the principal axes of the permittivity tensor. When referred to the principal axes, Eq. (39.5) assumes the form

$$D_x = \varepsilon_x E_x, \quad D_y = \varepsilon_y E_y, \quad D_z = \varepsilon_z E_z. \tag{39.16}$$

Since $\varepsilon_x \neq \varepsilon_y \neq \varepsilon_z$ in general, the vectors **D** and **E** are not collinear (Fig. 216).

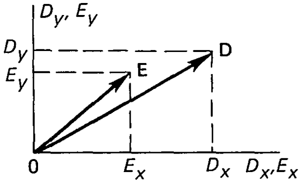


Fig. 216. Noncollinearity of vectors **D** and **E** in an anisotropic medium.

The anisotropy of magnetic properties of a medium is described analogously with the help of the permeability tensor. Henceforth, we shall assume the medium to be nonmagnetic and confine ourselves to description of the anisotropy of electric properties, since this is the situation most frequently encountered in problems on crystal optics. The magnetic anisotropy plays a significant role in the propagation of light in transparent ferrites and several other media.

?

What are the sources of anisotropy in electric properties of a medium?
How is the permittivity tensor reduced to its principal axes?
What is the relation between the electric displacement vector and the electric field strength in a Cartesian system of coordinates whose axes coincide with the principal axes of the permittivity tensor?
Which physical factor determines the symmetry of the permittivity tensor?

SEC. 40. PROPAGATION OF A PLANE ELECTROMAGNETIC WAVE IN AN ANISOTROPIC MEDIUM

Types of waves and their properties in an anisotropic medium are described.

PLANE ELECTROMAGNETIC WAVE IN ANISOTROPIC MEDIUM. In analogy with (2.50) and (2.51), the field vectors of a plane electromagnetic wave are presented in the form

$$\mathrm{E} = \mathrm{E}_0 e^{-i(\omega t - \mathrm{k}\cdot\mathrm{r})}, \quad \mathrm{D} = \mathrm{D}_0 e^{-i(\omega t - \mathrm{k}\cdot\mathrm{r})}, \quad \mathrm{B} = \mathrm{B}_0 e^{-i(\omega t - \mathrm{k}\cdot\mathrm{r})},$$
$$\mathrm{H} = \mathrm{H}_0 e^{-i(\omega t - \mathrm{k}\cdot\mathrm{r})}. \tag{40.1}$$

Substituting these relations into Maxwell's equations (2.1)-(2.5) and using (2.45) as well as the rules (2.52) for the operator ∇, we arrive at the following equations:

$$-\mathrm{k} \times \mathrm{H} = \omega \mathrm{D}, \tag{40.2a}$$
$$\mathrm{k} \times \mathrm{E} = \omega \mu_0 \mathrm{H}, \tag{40.2b}$$
$$\mathrm{k} \cdot \mathrm{D} = 0, \tag{40.2c}$$
$$\mathrm{k} \cdot \mathrm{H} = 0. \tag{40.2d}$$

The wave vector k is perpendicular to the constant-phase surface, and therefore indicates the direction of propagation of the wave front. The phase velocity v of the wave is directed along this vector; this direction is also considered as the direction of wave propagation and is characterized by the unit vector $\mathrm{n} = \mathrm{k}/k$. It can be seen from (40.2c) and (40.2d) that the wave propagates in a direction perpendicular to D and H. Expression (3.1) for Poynting's vector shows that the energy flux is perpendicular to E and H. The direction of energy flux in a wave is called the ray. Generally speaking, this direction is not the same as the direction of wave propagation. We denote the unit vector in the direction of the ray by $\tau = \mathrm{S}/S$. It was mentioned in Sec. 15 that the electromagnetic wave energy propagates with the group velocity. Hence it can be stated that the group velocity v_g has the same direction as τ.

Since vectors E and D are not collinear in an anisotropic medium, the directions of propagation of the wave and the ray do not coincide, and hence the group velocity and the phase velocity have also different directions. This is the first main feature in the propagation of an electromagnetic wave in an anisotropic medium. The second basic feature is that the velocity of propagation of electromagnetic waves depends on the direction of propagation and the polarization of the waves.

It can be seen from (40.2) and (3.1) that n and τ are perpendicular to H; E and D are also perpendicular to H. Moreover, n and τ are also perpendicular to D and E respectively. This means that D, E, n and τ lie in the same

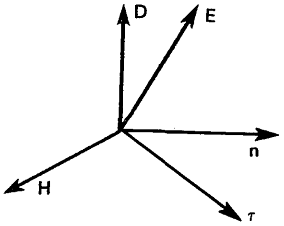


Fig. 217. Mutual arrangement of vectors of a plane wave in an anisotropic medium.

plane perpendicular to H (Fig. 217). The angle between D and E is equal to the angle between n and τ.

DEPENDENCE OF PHASE VELOCITY ON THE DIRECTION OF WAVE PROPAGATION AND OSCILLATIONS OF VECTOR D. Let us consider a wave propagating in the positive direction of the Z-axis, which is one of the principal axes of the permittivity tensor ε_{ij} (Fig. 215). We assume the electric displacement vector D to be collinear with the X-axis ($D_x \neq 0$, $D_y = 0$, $D_z = 0$), and hence the vector H is collinear with the Y-axis. On the basis of (39.16), we can write $E_x = D_x/\varepsilon_x$, $E_y = E_z = 0$. Equations (40.2a) and (40.2b) then assume the form

$$kH_y = \omega D_x, \tag{40.3a}$$

$$kE_x = \omega\mu_0 H_y. \tag{40.3b}$$

Termwise multiplication of the left- and right-hand sides of these equations gives

$$k^2 E_x H_y = \omega^2 \mu_0 \varepsilon_x H_y, \tag{40.4}$$

where $D_x = \varepsilon_x E_x$. Hence

$$k^2 = \omega^2 \varepsilon_x \mu_0, \tag{40.5}$$

and the phase velocity is therefore given by

$$\boxed{v_x = \omega/k = 1/\sqrt{\varepsilon_x \mu_0},} \tag{40.6}$$

where the subscript x on the phase velocity means that it is the velocity of a wave whose vectors D and E are collinear with the X-axis. If the vectors D and E are collinear with the Y-axis, we get, in analogy with (40.6),

$$\boxed{v_y = 1/\sqrt{\varepsilon_y \mu_0}.} \tag{40.7}$$

Since, in general $\varepsilon_x \neq \varepsilon_y$, we conclude that the phase velocity of the wave is different for these two directions of oscillations of the vector E. Hence only those waves can propagate along the Z-axis, whose vectors D and E oscillate parallel either to the X-axis or to the Y-axis.

!

Only two waves with mutually perpendicular linear polarizations and different velocities can propagate in a given direction in an anisotropic medium.

FRESNEL EQUATION. Similar conclusions can also be drawn concerning the waves propagating in the direction of the X-axis or the Y-axis. In order to study the behaviour of waves propagating in an arbitrary direction, we must analyze the general case instead of Eqs. (40.3a) and (40.3b), which are particular cases of Eqs. (40.2a) and (40.2b). Substituting the expression from (40.2b) into Eq. (40.2a) and using the formula for a vector triple product, we obtain

$$\underline{\overline{\mathrm{n}(\mathrm{n}\cdot\mathrm{E}) - \mathrm{E} + \mu_0 v^2 \mathrm{D} = 0,}} \tag{40.8}$$

where $v = \omega/k$. It is convenient to carry out all subsequent calculations in a Cartesian system of coordinates whose axes coincide with the principal axes of the permittivity tensor, when $D_i = \varepsilon_i E_i$. In this coordinate system, Eq. (40.8) can be written in the form of three scalar equations

$$n_i(\mathrm{n}\cdot\mathrm{E}) - E_i(1 - v^2\mu_0\varepsilon_i) = 0 \quad (i = 1, 2, 3). \tag{40.9}$$

Suppose that the vector E is directed along one of the principal axes of the permittivity tensor, say, the X-axis. In this case, $D_1 = \varepsilon_1 E_1$, $D_2 = D_3 = 0$ and $E_2 = E_3 = 0$. Obviously, the vectors τ and n coincide in this case and lie in the X_2X_3 plane (it should be recalled that $X_1 = X$, $X_2 = Y$ and $X_3 = Z$). Equation (40.9) can then be reduced to the form

$$E_1(1 - v_1^2\mu_0\varepsilon_1) = 0, \tag{40.10}$$

where the phase velocity is denoted by subscript 1 to indicate that the vectors E and D of the wave under consideration are collinear with the X-axis. Since $E_1 \neq 0$, we conclude from (40.10) that

$$v_1 = 1/\sqrt{\varepsilon_1\mu_0}. \tag{40.11}$$

This particular case coincides with the one considered as the simplest example derived from Eqs. (40.3a) and (40.3b), while formula (40.11) coincides with (40.6). Similar formulas are obtained for the X_2- and X_3-axes:

$$v_i = 1/\sqrt{\varepsilon_i\mu_0} \quad (i = 2, 3). \tag{40.12}$$

It can be stated that v_i is the phase velocity of the wave, corresponding to the X_i-axis. It should be noted, however, that it is not the projection of the phase velocity on the X_i-axis, but simply a characteristic of the phase velocity of a wave whose vectors E and D are collinear with the X_i-axis. The velocities v_i are called the principal velocities of wave propagation. Taking (40.12) into account, we can present Eqs. (40.9) in the form

$$n_i(\mathrm{n}\cdot\mathrm{E}) - E_i(1 - v^2/v_i^2) = 0. \tag{40.13}$$

Multiplying both sides of this equation by $n_i/(1 - v^2/v_i^2)$ and summing over i, we obtain the equality

$$\sum_{i=1} \frac{n_i^2 v_i^2}{v_i^2 - v^2} - 1 = 0, \tag{40.14}$$

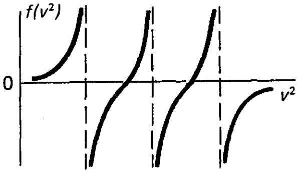


Fig. 218. To the graphic determination of the roots of Fresnel's equation.

where we have cancelled out the factor $\sum n_i E_i = \mathbf{n} \cdot \mathbf{E}$. Since $\sum n_i^2 = 1$, Eq. (40.14) assumes the following form after reduction to a common denominator and division by $v^2 \neq 0$:

$$\boxed{\sum_{i=1}^{3} \frac{n_i^2}{v_i^2 - v^2} = 0.} \tag{40.15}$$

This formula, called the Fresnel equation, allows us to determine the phase velocity in the direction characterized by the direction cosines n_1, n_2, and n_3. It should be observed that the quantities v_i in this equation are not the projections of the vector $\mathbf{n}v$ onto the coordinate axes, i.e. $n_i v \neq v_i$. The solution of Eq. (40.15) gives phase velocity v as a function of n_i and v_i.

It is convenient to express the phase velocity as a function of the direction of vector $\mathbf{D}$. Let $\mathbf{d} = \mathbf{D}/D$ be a unit vector in the direction of $\mathbf{D}$. A scalar multiplication of both sides of (40.8) by $\mathbf{D}$ gives

$$-\sum \varepsilon_i E_i^2 + \mu_0 v^2 D^2 = 0, \tag{40.16}$$

or

$$v^2 = (1/D^2) \sum_i v_i^2 D_i^2, \tag{40.17}$$

where $\varepsilon_1 E_i^2/\mu_0 = v_i^2 D_i^2$. Since $d_i = D_i/D$, Eq. (40.17) assumes the form

$$v^2 = \sum_i d_i^2 v_i^2, \tag{40.18}$$

i.e. the phase velocity is completely determined by the direction of vector $\mathbf{D}$.

POSSIBLE TYPES OF WAVES. Equation (40.15) will be satisfied if its left-hand side contains terms with opposite signs. Hence v^2 can be neither greater nor smaller than all v_i^2. To obtain the roots of Eq. (40.15), we plot the function

$$f(v^2) = \sum \frac{n_i^2}{v_i^2 - v^2} \tag{40.19}$$

This plot is shown in Fig. 218. The dashed vertical lines pass through points v_i^2. It can be seen from the figure that there are two real values of v_i^2 satisfying Eq. (40.15). This means that waves with two different phase velocities v' and v'' can propagate in a given direction. The velocities v' and v'' respectively lie between the minimum and mean values, and between mean and the maximum values of the velocities v_i. We shall prove that the vectors $\mathbf{D}'$ and $\mathbf{D}''$ of these waves are

?

In which cases do the electric displacement vector and the electric field strength of a wave in an anisotropic medium coincide? Why does the normal to the wave front surface in an anisotropic medium not coincide with the direction of energy flow of a wave in the general case? When do they coincide?

$\vec n(\vec n\cdot\vec E') - \vec E' + \mu_0 v'^2 \vec D' = 0$

$(\vec n\cdot\vec D'')(\vec n\cdot\vec E') - \vec E'\cdot\vec D'' + \mu_0 v'^2 \vec D'\cdot\vec D'' = 0$

$\Downarrow$

$-\vec E'\cdot\vec D'' + \mu_0 v'^2 \vec D'\cdot\vec D'' = 0$

и аналогично

$-\vec E''\cdot\vec D' + \mu_0 v''^2 \vec D'\cdot\vec D'' = 0$

mutually perpendicular. For this purpose, we form the scalar product of relation (40.8) for D′ and E′ with D″ and termwise subtract from the result the scalar product of the same relation (40.8) for D″ and E″ with D′. As a result, we obtain the relation

$$(\mathrm{D}'\cdot\mathrm{E}'' - \mathrm{D}''\cdot\mathrm{E}') - \mu_0(v'^2 - v''^2)\,\mathrm{D}'\cdot\mathrm{D}'' = 0. \tag{40.20}$$

Considering that

$$\mathrm{D}'\cdot\mathrm{E}'' = \sum_{i,j}\varepsilon_{ij}E'_jE''_i = \sum_{i,j}\varepsilon_{ji}E''_iE'_j = \mathrm{D}''\cdot\mathrm{E}', \tag{40.21}$$

we find that the difference in the first parentheses of (40.20) is equal to zero. Since $v'^2 - v''^2 \neq 0$, we conclude that

$$\boxed{\mathrm{D}'\cdot\mathrm{D}'' = 0,} \tag{40.22}$$

which means that the vectors D of two waves propagating in a given direction are mutually perpendicular. This means that in the plane perpendicular to the direction of propagation of each wave, there exist two mutually perpendicular directions with which the oscillations of the vector D are collinear. The phase velocities of the corresponding waves are different. There are no other waves propagating in the same direction but with their vector D oscillating in some other direction.

SEC. 41. PASSAGE OF RAYS THROUGH AN ANISOTROPIC MEDIUM

The passage of rays in an anisotropic medium is analyzed with the help of an ellipsoid of group velocities. The definition of uniaxial and biaxial crystals is given.

DEPENDENCE OF RAY (GROUP) VELOCITY ON DIRECTION. All the results concerning the direction of propagation of the wave front and the phase velocity were obtained by analyzing Eqs. (40.2) containing wave vector k and frequency ω, which characterize the phase velocity, as well as the normal n to the surface of the wave front. To study the problem of light rays and group velocity v_g, we must transform these equations in such a way that the formulas contain τ and v_g. In order to determine the group velocity v_g, it should be observed that the wave front propagates in the direction of n, while the energy propagates in the direction of τ. Hence the "front of energy flux" is perpendicular to τ. This leads to the conclusion that (see Fig. 217) the group and phase velocities of light in an anisotropic medium are connected through the relation

$$v_g = v\cos(\widehat{\mathrm{n},\tau}) \doteq v(\mathrm{n}\cdot\tau), \tag{41.1}$$

where $(\widehat{\mathrm{n},\tau})$ is the angle between n and τ.

$-\vec{k}\times H = \omega D$ (40,2a)

$\vec{k}\times\vec{E} = \omega\mu_0\vec{H}$ (40,2b)

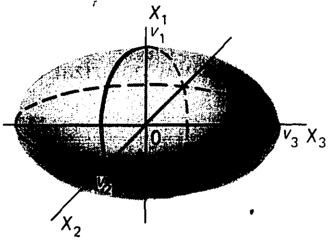


Fig. 219. Ellipsoid of group velocities.

Vector premultiplication of Eqs. (40.2a) and (40.2b) by τ gives

$$(\tau\cdot k)H = \omega\tau\times D, \tag{41.2a}$$

$$(\tau\cdot k)E = -\omega\mu_0\tau\times H. \tag{41.2b}$$

Substituting the expression for H from (41.2a) into (41.2b), we obtain

$$(\tau\cdot k)E = -\omega^2\mu_0[\tau(\tau\cdot D) - D]/(\tau\cdot k). \tag{41.3}$$

Taking into account the relations $k = nk$, $v = \omega/k$ and $v_g = v(n\cdot\tau)$, we can write Eq. (41.3) in the form

$$\tau(\tau\cdot D) - D + [1/(\mu_0 v_g^2)]E = 0. \tag{41.4}$$

This equation is identical to (40.8) if we make the following substitutions in the latter: $n\to\tau$, $E\to D$, $D\to E$, and $\mu_0 v^2\to 1/(\mu_0 v_g^2)$. Hence all the results obtained for a wave on the basis of (40.8) can be reformulated for rays by taking into account the above-mentioned substitutions.

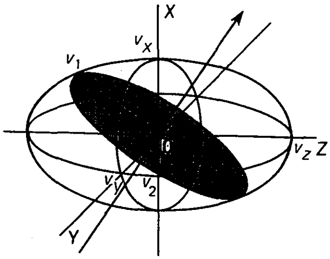


Fig. 220. To the analysis of the path of light rays through ellipsoid of group velocities.

Suppose that vector E is directed along any of the principal axes of the permittivity tensor (for example, the X-axis). Since in this case D is parallel to E, we obtain for v_g an expression which is identical to (40.10) or, in other words, the principal group velocities are identical to the principal phase velocities [see (40.12)]. Hence they can be denoted by v_i and not by v_{gi}. Instead of (40.13), we obtain

$$\tau_i(\tau\cdot D) - D_i(1 - v_i^2/v_g^2) = 0. \tag{41.5}$$

The Fresnel equation assumes the following form for ray velocities:

$$\sum_{i=1}^{3}\tau_i^2 v_i^2/(v_i^2 - v_g^2) = 0. \tag{41.6}$$

Denoting the unit vector in the direction of E by $I = E/E$, we can write the ray velocity equation corresponding to (40.18) in the following form:

$$1/v_g^2 = \sum_i l_i^2/v_i^2. \tag{41.7}$$

$$\cancel{(\vec{\tau}\cdot\vec{l})}(\vec{\tau}\cdot\vec{D}) - \vec{D}\cdot\vec{l} + \frac{\vec{E}\cdot\vec{l}}{\mu_0 v_g^2} = 0$$

0 E

$$\frac{1}{v_g^2} = \frac{\mu_0}{E}\vec{D}\cdot\vec{l} = \frac{\mu_0}{E}\sum_i D_i l_i =$$

$$= \frac{\mu_0}{E}\sum_i \varepsilon_0\varepsilon_i E_i l_i = \sum_i \mu_0\varepsilon_0\varepsilon_i\frac{E_i}{E}l_i$$

$\mu_0\varepsilon_0\varepsilon_i \to \frac{1}{v_i^2}$

$$= \sum_i \frac{l_i^2}{v_i^2}$$

We shall prove that the velocity in the direction of the ray is indeed the group velocity. In Eq. (40.15), we make the substitution $v = \omega/k$. This gives

$$\sum n_i^2/(\omega^2 - v_i^2 k^2) = 0. \tag{41.8}$$

Considering (41.8) as the equation which determines $\omega = \omega(k)$ in an implicit form and evaluating $v_{gi} = \partial\omega/\partial k_i$ from this equation, we can verify that these velocities satisfy Eq. (41.6). Thus, the velocities appearing in (41.6) are indeed group velocities.

Two waves propagating with two different group velocities in a given direction have mutually perpendicular polarizations:

$$\mathbf{E}'\cdot\mathbf{E}'' = 0. \tag{41.9}$$

This can be proved by following a procedure identical to the one used for deriving formula (40.22).

ELLIPSOID OF GROUP VELOCITIES. Making the substitution of variables

$$x_i \to \sqrt{\mu_0}\, x_i \tag{41.10}$$

in Eq. (39.14), we obtain

$$x_1^2/v_1^2 + x_2^2/v_2^2 + x_3^2/v_3^2 = 1, \tag{41.11}$$

where $v_1 = 1/\sqrt{\varepsilon_1\mu_0}$, $v_2 = 1/\sqrt{\varepsilon_2\mu_0}$ and $v_3 = 1/\sqrt{\varepsilon_3\mu_0}$. Equation (41.11) describes an ellipsoid (Fig. 219) identical to the ellipsoid described by Eq. (39.14) (see Fig. 215), but with different variables. The ellipsoid the points on whose surface satisfy Eq. (41.11) is called the ellipsoid of group velocities (x_i in this equation has the dimensions of velocity).

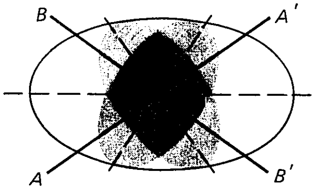


Fig. 221. Ellipsoid of group velocities of a biaxial crystal (*AA'* and *BB'* are the optical axes).

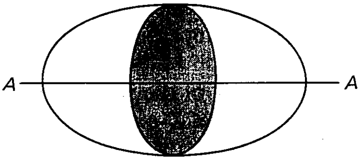


Fig. 222. Ellipsoid of group velocities of a uniaxial crystal (*AA'* is the optical axis).

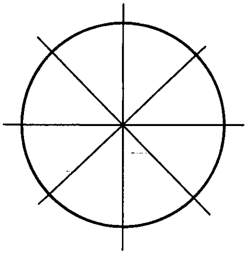


Fig. 223. Ellipsoid of group velocities of an optically isotropic medium.

!

The ellipsoid of group velocities gives a complete solution of the problem of light propagation in an anisotropic medium. For this purpose, use can also be made of the index ellipsoid and the ray surface.

ANALYSIS OF RAY PATH THROUGH ELLIPSOID OF GROUP VELOCITIES. The solution of Eqs. (41.5) and (41.6) can be obtained with the help of a geometrical construction based on the ellipsoid of group velocities (Fig. 220) described by the equation

$$x^2/v_x^2 + y^2/v_y^2 + z^2/v_z^2 = 1, \tag{41.12}$$

where v_x, v_y and v_z are the principal ray velocities.

The direction of the ray is specified by the unit vector τ. Through the centre of the ellipsoid of group velocities, we pass a plane perpendicular to τ. In the ellipsoid cross section, this plane forms an ellipse (Fig. 220) whose principal semiaxes are v_1 and v_2. Vector E of the light wave propagating along the ray can oscillate only in a direction parallel to the principal axes of this ellipse. The corresponding group (ray) velocities are equal to the lengths v_1 and v_2 of the principal semiaxes. Thus, only two linearly polarized waves with different group (ray) velocities can propagate in an arbitrary direction of τ. If the ellipsoid cross section degenerates into a circle, then $v_1 = v_2$, and hence any radius of the circle is a principal semiaxis. In this case, vector E can oscillate in any direction perpendicular to τ.

OPTICAL AXIS. In the direction perpendicular to the circular cross section of an ellipsoid of group velocities, all rays have the same ray velocity and vectors E of the waves can oscillate in any direction of the plane of circular cross section.

This means that the medium anisotropy is not manifested for such rays, and the medium behaves as an isotropic medium. The direction perpendicular to the circular cross section is called the optical axis of the anisotropic medium.

BIAXIAL AND UNIAXIAL CRYSTALS. In the analytical geometry, it is proved that an ellipsoid with three different values of the principal axes has two circular cross sections (Fig. 221). This means that if all the principal velocities v_x, v_y and v_z are different for an ellipsoid of group velocities, the medium will have two optical axes, viz. AA' and BB'. Anisotropy is usually observed in crystals, and hence it is customary to speak of the optical axes of a crystal. Crystals with two optical axes are called biaxial crystals.

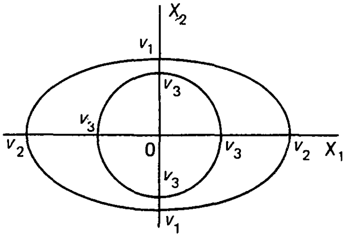


Fig. 224. Cross section of the ray surface by the plane $X_3 = 0$ for $v_3 < v_1 < v_2$.

If two of the principal velocities in an ellipsoid of group velocities are equal, the latter becomes an ellipsoid of revolution about the third axis. In this case, we have only one optical axis which coincides with the axis of revolution (i.e. with the third principal axis of the ellipsoid). Such crystals are called uniaxial crystals (Fig. 222).

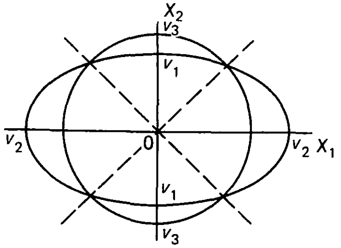


Fig. 225. Cross section of the ray surface by the plane $X_3 = 0$ for $v_1 < v_3 < v_2$.

If all the principal velocities of an ellipsoid of group velocities are equal, the ellipsoid is transformed into a sphere. In this case, all directions are identical, and the crystal is said to be optically isotropic (Fig. 223).

INDEX ELLIPSOID. Instead of rays, we can consider the propagation of waves whose motion is characterized by a unit vector **n** normal to the surface of the wave front. The analysis of the ray paths is carried out by using vectors **E** and **H** which are perpendicular to the direction of ray τ. The wave propagation is analyzed with the help of vectors **D** and **H** which are perpendicular to vector **n**. Instead of the ellipsoid of group velocities, we must now use the index ellipsoid whose equation is obtained from the quadratic form (39.8) in the same way as for the ellipsoid of group velocities. In accordance with (39.12), the quantities (x, y, z) in formula (39.14) are proportional to E_x, E_y, and E_z. In order to take D_x, D_y, and D_z as independent variables, we must make the substitution $x \to x/\varepsilon_x$, $y \to y/\varepsilon_y$ and $z \to z/\varepsilon_z$. As a result, we obtain the following relation instead of (39.14):

$$x^2/\varepsilon_x + y^2/\varepsilon_y + z^2/\varepsilon_z = 1. \tag{41.13}$$

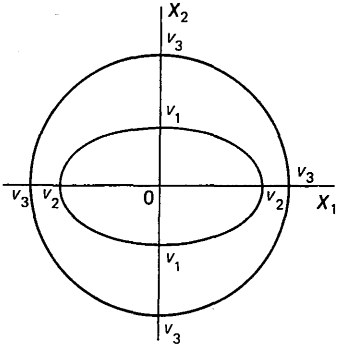


Fig. 226. Cross section of the ray surface by the plane $X_3 = 0$ for $v_1 < v_2 < v_3$.

In the same way as (39.14) was used to obtain equation (41.12) for the ellipsoid of group velocities, we use (41.13) to derive the relation

$$v_x^2 x^2 + v_y^2 y^2 + v_z^2 z_z^2 = 1, \tag{41.14}$$

called the index ellipsoid equation. The values of the velocities v_x, v_y and v_z here are the same as in (39.14). In order to analyze wave propagation with the help of this ellipsoid, it is convenient to present Eq. (41.14) in the form

$$\frac{x^2}{(1/v_x)^2} + \frac{y^2}{(1/v_y)^2} + \frac{z^2}{(1/v_z)^2} = 1. \tag{41.15}$$

The analysis of the wave propagation is carried out in the same way as for the rays, with the only difference that the index ellipsoid is used instead of the ellipsoid of group velocities. The direction of the wave propagation is given by vector n. The cross section of ellipsoid (41.15) by a plane perpendicular to n and passing through the centre of the ellipsoid is determined. Oscillations of vector D are possible only in the directions parallel to the principal axes of the ellipse in the cross section of the ellipsoid. The phase velocities of the waves are inversely proportional to the lengths of the corresponding principal axes of the ellipse. For analyzing light propagation in anisotropic media, however, it is more convenient to use the concept of ray surface rather than the surface of the wave front.

RAY SURFACE. In an anisotropic medium, rays can also be analyzed directly with the help of Fresnel equation (41.6) without even considering the ellipsoid of group velocities. For this purpose, we go over to new variables

$$\mathrm{r} = \tau v_g, \quad r_i = x_i = \tau_i v_g. \tag{41.16}$$

In terms of these variables, Eq. (41.6) assumes the form

$$\sum_i x_i^2 v_i^2/(r^2 - v_i^2) = 0. \tag{41.17}$$

The fourth-order surface defined by this equation is called the ray surface. The distance r from the origin of coordinates to the corresponding point on the surface is proportional to the ray velocity in the direction of τ. In each direction, the ray surface is encountered twice. This corresponds to the existence of two velocities of light propagation in each direction.

Let us consider the section of the ray surface by coordinate planes, say, by the plane $x_3 = 0$. Under this condition, Eq. (41.17) is split into two equations:

$$x_1 v_1^2 (r^2 - v_1^2) + x_2^2 v_2^2/(r^2 - v_2^2) = 0, \; (r^2 = x_1^2 + x_2^2), \tag{41.18a}$$

$$r^2 - v_3^2 = x_1^2 + x_2^2 - v_3^2 = 0. \tag{41.18b}$$

After some elementary transformations, the first equation assumes the form

$$x_1^2/v_2^2 + x_2^2/v_1^2 = 1. \tag{41.19}$$

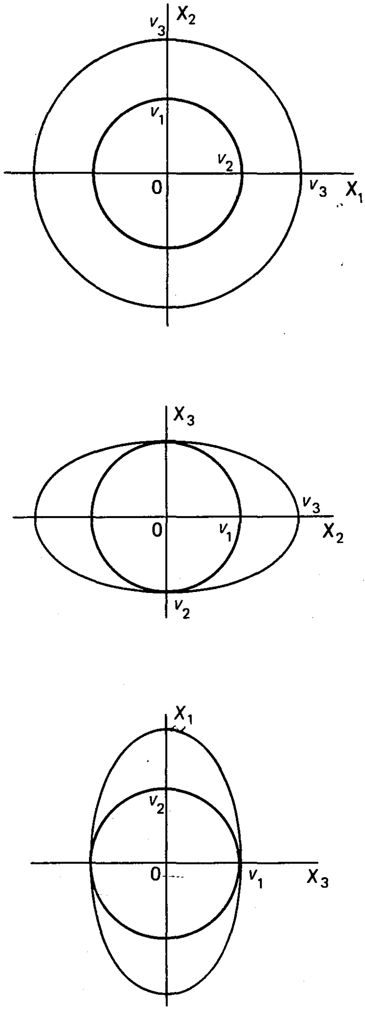


Fig. 227. Cross sections of the ray surface of a uniaxial crystal for $v_3 > v_1 = v_2$.

?

What is an optical axis? How many optical axes may exist in a crystal? What are uniaxial and biaxial crystals? Describe the method of analysis of ray propagation in an anisotropic medium with the help of the ellipsoid of group velocities.

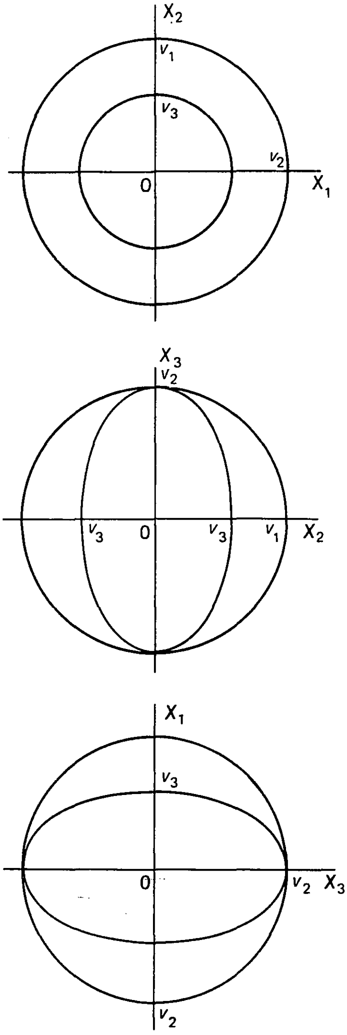


Fig. 228. Cross sections of the ray surface of a uniaxial crystal for $v_3 < v_1 = v_2$.

Thus, the ray surface is intersected by the coordinate planes along ellipse (41.19) and circle (41.18b). Their mutual arrangement is determined by the relation between the velocities v_1, v_2 and v_3. For $v_3 < v_1 < v_2$, the cross sections have the form shown in Fig. 224. If $v_3 < v_2 < v_1$, the ellipse is elongated along the X_2-axis. The cross sections for $v_1 < v_3 < v_2$ are shown in Fig. 225. For $v_2 < v_3 < v_1$, the ellipse is elongated along the X_2-axis. The cross sections for $v_1 < v_2 < v_3$ are presented in Fig. 226. For $v_2 < v_1 < v_3$, the ellipse is elongated along the X_2-axis.

It can be easily seen that the circle describes the velocity of the ray whose electric vector oscillates along the principal axis perpendicular to the plane under consideration (along the principal axis X_3 in the present case), i.e. the electric vector oscillates in a direction perpendicular to the plane of the figure. The electric vector of the ray described by the ellipse oscillates in the plane of the figure (in the present case, in the X_1X_2-plane). Figures 224-226 show the possible cross sections of the ray surface by the coordinate planes for unequal values of v_1, v_2 and v_3. Since the optical axis is determined by the fact that the velocities for both rays in the direction of the axis are equal, the axis can be found by using the geometric construction shown in Fig. 225. Here, the optical axes are shown by the dashed lines. For unequal values of v_1, v_2 and v_3, the crystal has two optical axes.

For uniaxial crystals, the two axes of the ellipsoid of group velocities are equal. We put $v_1 = v_2$. Then for $v_3 > v_1 = v_2$, the ellipsoid of group velocities is oblate along the X_3-axis, while for $v_3 < v_1 = v_2$, the ellipsoid is oblong along the X_3-axis. The sections of the ray surface by the coordinate planes in these cases are shown in Figs. 227 and 228. The optical axis coincides with the principal axis of the ellipsoid of group velocities. The crystals for which $v_3 < v_1 = v_2$ are called positive crystals, while crystals for which $v_3 > v_1 = v_2$ are called negative crystals.

The ray surfaces can also be constructed without resorting to the solution of (41.17), by proceeding directly from the ellipsoid of group velocities. For this purpose, we lay off two segments from the centre of the ellipsoid in each direction, equal to the principal axes of the ellipses in the ellipsoid cross sections perpendicular to these directions. These segments are equal to the ray velocities. The ends of the segments lie on the ray surface. Such a method, based on the ellipsoid of ray velocities, is quite visual and is preferable for a qualitative analysis of the ray surfaces. For a quantitative analysis, it is more expedient to use Eq. (41.17).

SEC. 42. BIREFRINGENCE

Birefringence and the phenomena related to it are considered with the help of the ray surfaces.

ORDINARY AND EXTRAORDINARY RAYS. The plane passing through a ray inclined to the optical axis (Fig. 229) and the optical axis is called the principal plane (in Fig. 229 this plane coincides with the plane of the figure). Obviously, for a ray whose vector E_0 is perpendicular to the principal plane, the velocity is independent of the direction and is equal to the ray velocity collinear with the optical axis. This ray is called ordinary and the quantities corresponding to it are assigned a subscript "o" (for example, its velocity is designated by v_o and the refractive index for it is denoted by $n_o = c/v_o$. For a ray whose vector E_e (Fig. 229) lies in the principal plane, the velocity depends on the direction, since the principal axis of the ellipse formed as a section of the ellipsoid varies with the direction of the ray. Such a ray is called extraordinary, and the quantities associated with it are assigned a subscript "e". Its velocity is denoted by v_e and the refractive index by n_e.

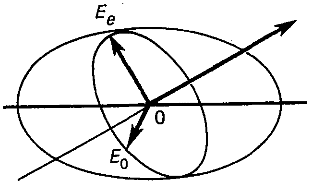


Fig. 229. Ordinary and extraordinary rays in a negative crystal.

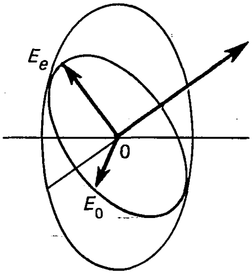


Fig. 230. Ordinary and extraordinary rays in a positive crystal.

For negative crystals, $v_o < v_e$, while for positive crystals $v_o > v_e$ (see Sec. 41). The ratio of the principal axes of an ellipsoid of group velocities is shown in Fig. 229 for a negative crystal and in Fig. 230 for a positive crystal. The cross sections of ray surfaces for negative crystals are shown in Fig. 227 and those for a positive crystal, in Fig. 228.

ESSENCE OF BIREFRINGENCE. Since only two rays with different velocities can propagate inside a crystal, the refraction of a ray at the surface of a crystal leads to the emergence of two rays in the crystal. The splitting of a ray into two by a crystal is called **birefringence.**

Birefringence was discovered in 1669 by Bartholin (1625–1698). It was first explained by Huygens who used the concept of an elliptical secondary wave.

HUYGENS' CONSTRUCTION. Birefringence is analyzed with the help of a construction proposed by Huygens. It is a generalization of Huygens' construction for isotropic media, used for the derivation of the Snell law (Fig. 231), and is carried out with the help of ray (not wave!) surfaces. For a positive crystal, whose optical axis is inclined at an arbitrary angle to the crystal surface, the Huygens construction is shown in Fig. 232. Segment AB is taken equal to unity. Point O is assumed to be the centre of the cross section of the ray surface. The radius of the circular cross section for an ordinary ray is equal to $1/n_o$, while the elliptic cross section for an extraordinary ray is plotted in such a way that the distance from the centre to a point on the ellipse is $1/n_e = v_e/c$. After this tangents are drawn from B to the circle and the ellipse. The straight lines passing through O and

!

For the ordinary ray, the electric field strength vector is perpendicular to the principal plane, while for the extraordinary ray the vector lies in the principal plane.

the points of tangency are the required rays: the ordinary ray passes through the point of tangency on the circle, and the extraordinary one through the point of tangency on the ellipse.

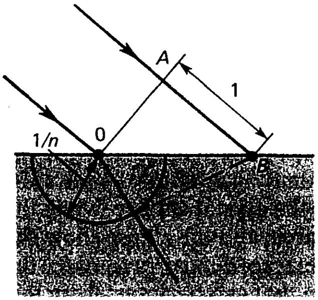


Fig. 231. Huygens' construction for isotropic media.

Figure 232 shows the simplest situation when the ray incident on a crystal lies in its principal plane. This allows all constructions to be made in the plane of the figure. If the ray entering the crystal does not lie in the principal plane of the crystal, the Huygens construction becomes three-dimensional. In this case, we must construct ellipsoids, spheres and planes, but the principle of finding the refracted rays remains the same and the refracted rays pass from point O through the points of contact of the ellipsoid and the sphere with the corresponding planes.

Let us consider various typical cases of birefringence.

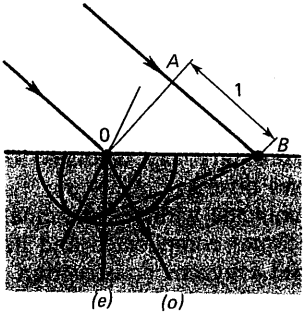


Fig. 232. Huygens' construction for a positive crystal in the case of an arbitrary angle between the optical axis and the crystal surface.

OPTICAL AXIS IS PERPENDICULAR TO THE CRYSTAL SURFACE. In the case of normal incidence, the ray is parallel to the optical axis and hence propagates in the same way as in an isotropic medium, i.e. there is no birefringence. Birefringence is observed in the case of an oblique incidence of the ray on the crystal surface. The nature of the birefringence depends on the type of the crystal. In a negative crystal (Fig. 233), the ordinary ray is refracted more strongly than the extraordinary ray. The points and arrows in this figure indicate the direcion of oscillation of the electric vector of the wave. In a positive crystal (Fig. 234), it is the extraordinary ray that is refracted more strongly.

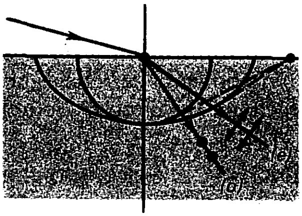


Fig. 233. Birefringence at the surface of a negative crystal when the optical axis is perpendicular to the crystal surface.

OPTICAL AXIS IS PARALLEL TO THE CRYSTAL SURFACE. For a normally incident ray (Fig. 235), two rays are formed in the crystal. In the ordinary ray, vector E is perpendicular to the optical axis (this direction is indicated by points in Fig. 235), while for the extraordinary ray, the vector is parallel to the optical axis (shown by ordinary arrows in Fig. 235). The rays coming out of a crystal plate acquire a phase difference and form an elliptically polarized wave as a result of superposition (see Sec. 43). If the incident light is natural, elliptically polarized waves with all types of orientations of the ellipses and ratios of their axes are formed at the exit from the crystal. This light is also natural in its properties.

When light is incident at an angle to the surface of the crystal, the nature of birefringence depends on the angle between the plane of incidence and the principal plane. If the incidence plane and the principal plane coincide (Fig. 236), the ordinary and extraordinary rays lie in the same plane. Figure 236 presents birefringence for a negative crystal. The Huygens construction for a positive crystal is much the same.

If the incidence plane is perpendicular to the optical axis (Fig. 237), the ordinary and extraordinary rays lie in the incidence plane, but the refractive index for both rays is

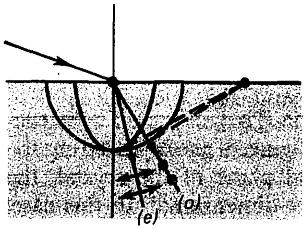


Fig. 234. Birefringence at the surface of a positive crystal when the optical axis is perpendicular to the crystal surface.

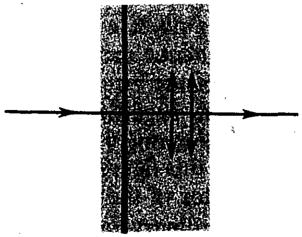


Fig. 235. Birefringence in the case of normal incidence at the crystal surface, when the optical axis is parallel to the surface.

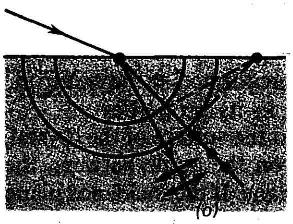


Fig. 236. Birefringence at the surface of a negative crystal when the optical axis is parallel to the surface and the plane of incidence coincides with the principal plane.

independent of the direction. The optical axis in Fig. 237 is perpendicular to the plane of the paper.

If the incidence plane intersects the optical axis at an angle other than $\pi/2$, the birefringence pattern becomes complicated since we have to draw a three-dimensional Huygens construction in this case. Figure 238 shows the section of ray surfaces by the crystal surface. This section coincides with the plane of the figure. The arrow terminating at point O shows the projection of the incident ray onto the crystal surface, while the dashed arrows terminating at the circle and the ellipse show the projections of the ordinary and extraordinary rays onto the crystal surface. The main conclusion that can be drawn from this situation is that the ordinary ray lies on the incidence plane, while the extraordinary ray does not. To obtain a more detailed graphic information, it is necessary to construct the three-dimensional model.

OPTICAL AXIS IS AT AN ANGLE TO THE CRYSTAL SURFACE. The simplest case in which the incidence plane coincides with the principal plane is shown in Fig. 232 to explain the essence of the Huygens construction. For normal incidence (Fig. 239), the ordinary ray has the same direction as the incident ray, while the extraordinary ray changes its direction. As a result, the two waves diverge.

If the incidence plane does not coincide with the principal plane, the birefringence pattern becomes complicated and its analysis requires a three-dimensional Huygens construction.

MALUS' LAW OF RAYS. To analyze the normal incidence of a ray on a plate cut out of a crystal parallel to the optical axis (Fig. 235), we must first determine the oscillation amplitudes in ordinary and extraordinary rays. This problem is solved with the help of Malus' law. If β is the angle between the optical axis and the line of oscillation of vector E (Fig. 240), and I is the intensity of the incident ray, the intensity of ordinary and extraordinary rays is given by

$$I_o = I \sin^2 \beta \quad \text{and} \quad I_e = I \cos^2 \beta . \tag{42.1}$$

This means that the electric vector E of the incident wave can be presented as the sum of the components parallel and perpendicular to the optical axis, which are the electric vectors of the ordinary and extraordinary rays.

POLARIZATION IN BIREFRINGENCE. Since ordinary and extraordinary rays are linearly polarized in mutually perpendicular planes, birefringence may be used for obtaining polarized rays. For this purpose, we must isolate ordinary and extraordinary rays in space, or eliminate one of these through strong absorption.

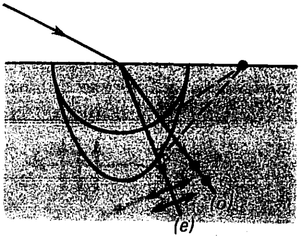


Fig. 237. Birefringence in the case when the plane of incidence is perpendicular to the optical axis.

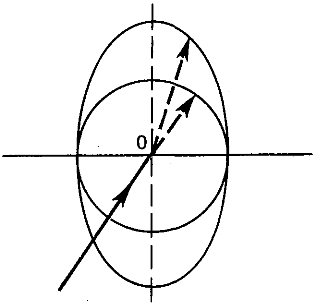


Fig. 238. Projection of rays onto the crystal surface when the optical axis is not perpendicular to the plane of incidence.

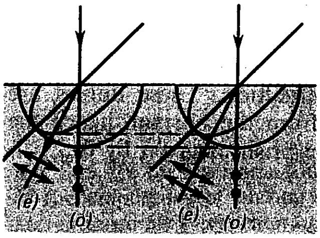


Fig. 239. Birefringence in the case of normal incidence when the optical axis of the crystal is not parallel to the surface.

POLAROID. If one of the rays emerging from a crystal plate (see Fig. 235) is strongly attenuated as a result of absorption, the light coming out of the plate will be linearly polarized. Such a plate is called a polaroid. Tourmaline crystals are good polaroids. In tourmaline crystals just 1 mm thick, ordinary rays are almost completely absorbed. In the transmitted (extraordinary) ray, the electric vector oscillates parallel to the optical axis (see Fig. 235). Herapathite is another good polaroid. With a thickness of just 0.1 mm, it absorbs one of the rays almost completely.

If a polaroid is employed for obtaining polarized light, it is called a polarizer. If, on the other hand, it is used for analyzing light polarization, it is called an analyzer (see Sec. 43).

POLARIZING PRISMS AND BIREFRINGENT PRISMS. Combinations of crystals producing polarized light are called polarizing or birefringent prisms. A polarizing prism is one having one polarized ray at the exit, while a prism from which both rays emerge is a birefringent prism.

NICOL PRISM. This is a polarizing prism made of Iceland spar. The crystals are cut relative to the optical axis as shown in Fig. 241, and are joined together with Canada balsam along the surface marked by a darker line in the figure. The refractive index of Canada balsam is $n = 1.550$, and lies between the refractive indices for ordinary and extraordinary rays. For an appropriate choice of the direction of the incident ray, the extraordinary ray passes through the prism while the ordinary ray suffers the total reflection at the Canada balsam layer and leaves the prism or is absorbed by its darkened face. The Nicol prism (it is frequently called just Nicol) is the most widely used polarizing prism.

BIREFRINGENT PRISM. A birefringent prism formed by glass and Iceland spar is shown in Fig. 242. The direction of the optical axis of the crystal is shown by dots. Birefringent prisms formed by combinations of prisms of Iceland spar with different mutual orientations of the optical axes are shown in Figs. 243*a*, *b*, *c*.

POLYCHROISM. The absorption of ordinary and extraordinary rays in a crystal depends on the wavelength as well as the direction of the incident wave. Consequently, the coloring of a crystal, through which white light is passing, depends on the direction of light propagation. This phenomenon is called polychroism.

Example 42.1. Figure 243 shows a Wollaston polarizing prism made of Iceland spar. The refractive index for the ordinary ray is $n_0 = 1.658$, while the refractive index for the extraordinary ray, when vector E is collinear with the optical

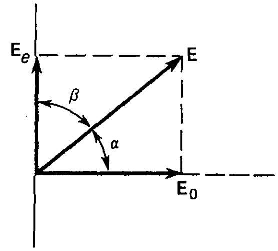


Fig. 240. Interpretation of Malus' law.

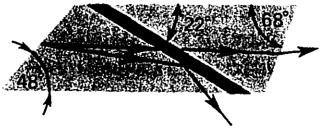


Fig. 241. Nicol prism.

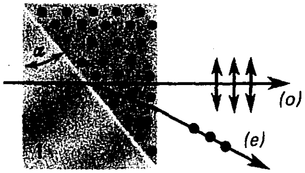


Fig. 242. A birefringent prism made of glass and Iceland spar.

axis, has the value $n_e = 1.486$. The angle $\alpha = 15°$. Find the angle between the rays emerging from the prism.

Upon passing the interface between media with mutually perpendicular optical axes, an ordinary ray in the first medium becomes an extraordinary ray in the second medium. Denoting the angles of refraction of ordinary and extraordinary rays by $\theta_{\mathrm{refr}\,o}$ and $\theta_{\mathrm{refr}\,e}$, we can write

$$\frac{\sin\theta_{\mathrm{in}\,o}}{\sin\theta_{\mathrm{refr}\,e}} = \frac{n_e}{n_o} < 1, \quad \frac{\sin\theta_{\mathrm{in}\,e}}{\sin\theta_{\mathrm{refr}\,o}} = \frac{n_o}{n_e} > 1. \tag{42.2}$$

Here $\sin\theta_{\mathrm{in}\,o} = \sin\theta_{\mathrm{refr}\,e} = \sin\alpha$.

Denoting the angles of refraction for the rays emerging from the prism in the air by $\theta'_{\mathrm{refr}\,o}$ and $\theta'_{\mathrm{refr}\,e}$, we can write the laws of refraction in the form

$$\begin{aligned} &\sin(\alpha - \theta_{\mathrm{refr}\,e})/\sin\theta'_{\mathrm{refr}\,e} = 1/n_e, \\ &\sin(\alpha - \theta_{\mathrm{refr}\,o})/\sin\theta'_{\mathrm{refr}\,o} = 1/n_o, \end{aligned} \tag{42.3}$$

since the angles of incidence at the boundary where the ordinary and extraordinary rays leave the prism are equal to $\alpha - \theta_{\mathrm{refr}\,o}$ and $\alpha - \theta_{\mathrm{refr}\,e}$ respectively. Taking (42.2) into account, we obtain from (42.3) the values $\theta'_{\mathrm{refr}\,e} = 2°14'$ and $\theta_{\mathrm{refr}\,o} = 3°2'$ for the angles of refraction. These angles are measured in opposite directions from the normal, hence the angle between the rays leaving the prism is given by $\theta'_{\mathrm{refr}\,o} + \theta_{\mathrm{refr}\,e} = 5°16'$.

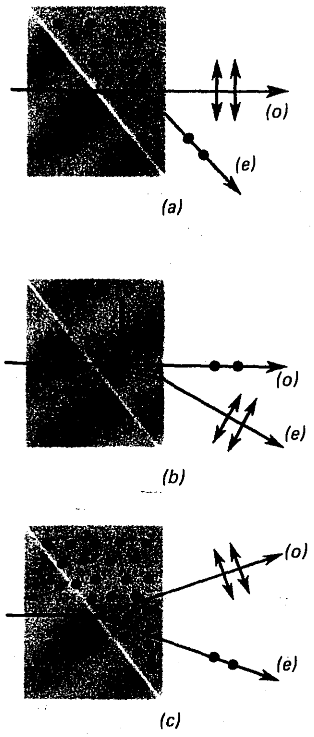


Fig. 243. A birefringent prism made of Iceland spar.

SEC 43. INTERFERENCE OF POLARIZED WAVES

Interference of waves with mutually perpendicular planes of polarization is described.

INTERFERENCE OF RAYS WITH MUTUALLY PERPENDICULAR DIRECTIONS OF LINEAR POLARIZATION. Rays with mutually perpendicular linear polarizations can be obtained as shown in Fig. 235 if a Nicol prism is placed in the path of the rays before they enter a crystal plate (Fig. 244). The Nicol prism produces a linearly polarized light, and the direction of electric vector oscillations can be varied by turning the prism. Linearly polarized light is incident on the crystal where it is split into the ordinary and extraordinary rays. In accordance with Malus' law of rays, we can write for the amplitude of these rays (Fig. 245)

$$a = A\cos\alpha, \qquad b = A\sin\alpha, \tag{43.1}$$

where A is the amplitude of the incident ray. The oscillations of the electric vector of the ordinary ray (along the X-axis) and the extraordinary wave at the entrance to the crystal can be presented in the form

$$x = a\cos\omega t', \qquad y = b\cos\omega t', \tag{43.2}$$

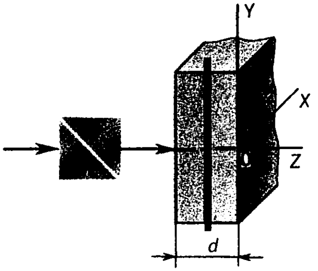


Fig. 244. Interference of rays with mutually perpendicular directions of linear polarization.

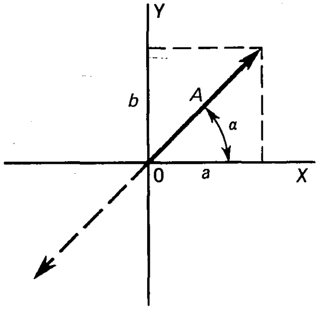


Fig. 245. To the determination of the amplitudes of ordinary and extraordinary rays entering a plate.

where t' is the time at which the oscillations at the entrance to the crystal are described. The ordinary ray spends a time $n_o d/c$ in passing through a plate of thickness d, while the corresponding value for the extraordinary ray is $n_e d/c$. Denoting by t the running time at which the rays emerge from the crystal, we obtain the following relations between the entrance time and the exit time for the ordinary and extraordinary rays:

$$t = t' + n_o d/c, \qquad t = t' + n_e d/c. \tag{43.3}$$

Hence, on the basis of (43.2) and (43.3), we can describe the electric vector oscillations of the ordinary and extraordinary waves leaving the crystal as follows:

$$x = a\cos(\omega t - kn_o d), \qquad y = b\cos(\omega t - kn_e d), \tag{43.4}$$

where $k = \omega/c$. Thus, a phase difference $\varphi = kd(n_e - n_o)$ is created between the waves. Relations (43.4) can be presented in the form

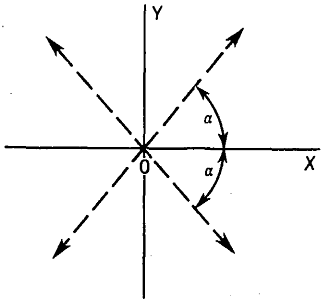


Fig. 246. Rotation of the plane of oscillations at the exit from a half-wave plate.

$$x = a\cos\xi,\qquad x = b\cos(\xi - \varphi), \tag{43.5}$$

where $\xi = \omega t - kn_o d$. These equatons can be considered as the parametric representation of an ellipse whose equation has the form (see Sec. 5)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{2xy}{ab}\cos\varphi = \sin^2\varphi. \tag{43.6}$$

The shape and orientation of the ellipse depend on α and φ (α depends on the ratio a/b).

QUARTER-WAVE PLATE. For $d(n_e - n_o) = \pm\lambda/4$, $\varphi = \pm\pi/2$, Eq. (43.6) assumes the form

$$x^2/a^2 + y^2/b^2 = 1. \tag{43.7}$$

This means that the principal axes of the ellipse coincide with the principal axes of the crystal. For $a = b$, i.e. for $\alpha = \pi/4$, Eq. (43.7) is transformed into the equation of a circle:

$$x^2 + y^2 = a^2. \tag{43.8}$$

In other words, a circularly polarized wave emerges from the crystal.

The sign of the difference $n_e - n_o$ depends on the type of the crystal. For a positive crystal $(n_e - n_o) > 0$, while for a negative crystal $(n_e - n_o) < 0$. The direction of revolution of vector E depends not only on the type of crystal, but also on the plate thickness d. This is so because Eq. (43.6) assumes the form of (43.7) not only for $d(n_e - n_o) = \pm\lambda/4$, but also for $d(n_e - n_o) = \lambda/4 + m\lambda/2$, where $m = \pm 1, \pm 2, \dots$.

HALF-WAVE PLATE. For $d(n_e - n_o) = \pm\lambda/2$, $\varphi = \pm\pi$, Eq. (43.6) assumes the form

$$(x/a + y/b)^2 = 0,\quad x/a + y/b = 0. \tag{43.9}$$

Hence at the exit from the plate we get a linearly polarized wave for which the quadrants of the XY-plane in which oscillations occur are different (Fig. 246), i.e. the polarization plane is turned through an angle 2α. Naturally, such a situation is possible not only when $d(n_e - n_o) = \pm\lambda/2$, but also when $d(n_e - n_o) = \lambda/2 + m\lambda$, where $m = \pm 1, \pm 2, \dots$.

WAVE PLATE. For $(n_e - n_o)\,d = m\lambda$, $\varphi = 2\pi m$, Eq. (43.6) assumes the form

$$(x/a - y/b)^2 = 0,\quad x/a - y/b = 0. \tag{43.10}$$

This means that the linearly polarized wave emerging from the crystal has the same plane of oscillations of the vector E as the wave entering the crystal.

?

In what case does the refracted ray lie in the plane of incidence?

When does the refracted ray emerge from the plane of incidence?

List polarizing and birefringence prisms that you know.

For what purposes are polaroids used?

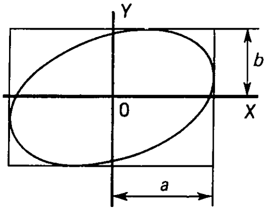


Fig. 247. To the analysis of elliptically polarized light.

ANALYSIS OF LINEARLY POLARIZED LIGHT. A Nicol prism is placed in the path of a ray in such a way that its principal direction is perpendicular to the direction of the ray. Two outcomes are possible when this direction is changed (by turning the Nicol prism): the intensity of the light transmitted through the Nicol prism changes, or it remains the same. If the intensity of the transmitted light remains unaltered, the light is either nonpolarized or polarized circularly. (The method of distinguishing between nonpolarized and circularly polarized light will be described below.) If the intensity of the transmitted light varies as a result of turning the prism, two situations are possible: either the intensity drops to zero, or it never becomes equal to zero. If the intensity decreases to zero, the light is linearly polarized, and the electric vector oscillations are collinear with the principal axis of the Nicol in the position when the intensity of the light transmitted through it is maximum. If the intensity does not drop to zero, then light is either elliptically polarized or partially polarized. The reason behind the change in the intensity of the light transmitted through the Nicol when the latter is turned for elliptically polarized light is shown in Fig. 247. The amplitude of the oscillations transmitted through the Nicol is equal to the distance between the tangents to the ellipse that are perpendicular to the direction of the transmitted oscillations. It can be seen that the maximum transmission amplitude is equal to the semi-major axis of the ellipse, while the minimum amplitude is equal to the semi-minor axis.

ANALYSIS OF ELLIPTICALLY POLARIZED LIGHT. An elliptically polarized wave can be treated as a superposition of two linearly polarized waves with a phase difference $\pi/2$ under the condition that the directions of their polarization coincide with the principal axes of the ellipse. Hence if a quarter-wave plate is placed in the path of an elliptically polarized ray in such a way that the principal axes of the plate and the ellipse coincide, a linearly polarized wave will emerge from the plate. The principal axes of the ellipse are determined by orienting the Nicol prism to positions where it transmits light with maximum and minimum intensities. Hence, by aligning the principal axis of a quarter-wave plate with one of the directions determined in this way and analyzing the light emerging from the plate for linear polarization, we can conclude whether the light is elliptically polarized or partially polarized. If the light emerging from the quarter-wave plate is linearly polarized, the light entering it is elliptically polarized. In the case of a partially polarized light emerging from the quarter-wave plate, the light entering it is also partially polarized.

!

The polarization of the wave formed as a result of interference of two waves with mutually perpendicular linear polarizations depends on the phase difference between them and the ratio of their amplitudes.

?

Describe the method of determining the polarization of light.
What is the difference between partially and elliptically polarized light?
What is the reason behind the formation of a cross when interference is observed in convergent rays?

ANALYSIS OF CIRCULARLY POLARIZED LIGHT. If the intensity of light emerging from a Nicol prism does not change as a result of rotating the prism, the light entering the prism is either circularly polarized or nonpolarized. We place a quarter-wave plate in the path of the light, but unlike the case of elliptic polarization, there is no need to bother about the choice of the direction of the principal optical axis of the plate. If the light emerging from the quarter-wave plate is linearly polarized, the light entering it is circularly polarized. If the transmitted light is nonpolarized, so is the light entering the plate.

Fig. 248. A compensator.

COMPENSATORS. The analysis of the elliptically polarized light can be simplified with the help of a **compensator** which is a plate along which the phase difference between the rays emerging from it is continuously varying. For the compensator shown in Fig. 248, we have

$$\Delta = (n_e d_1 + n_o d_2) - (n_o d_1 + n_e d_2) = (n_e - n_o)(d_1 - d_2). \quad (43.11)$$

Hence all possible phase differences arise between the rays leaving the plate. Linear polarization occurs at places where the phase difference between the rays is 0, 2π, 4π, … . In all other cases the emerging light is elliptically (or circularly) polarized with varying parameters of the ellipse. Observing the compensator surface through a Nicol prism, we find that illumination intensity varies periodically, indicating that the light entering the compensator is polarized. The phase difference between the rays created by the plate is known [see (43.11)]. Hence, knowing the intensity distribution for various positions of the polarizer (analyzer), we can determine the nature of polarization of the light entering the plate.

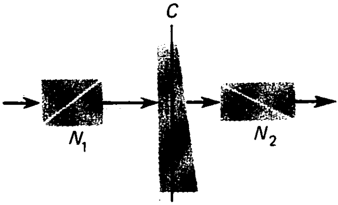


Fig. 249. Explaining the emergence of colours in thin films.

COLOURS OF CRYSTAL PLATES. Let us consider a crystal plate of varying thickness, its optical axis being perpendicular to the ray passing through it (Fig. 249). Nicol prism N_1 transmits only those oscillations of **E** that are parallel to its principal direction. On account of the varying thickness of plate C, the light emerging from it is elliptically polarized with different orientations of the ellipses. Analyzer N_2 retains from each oscillation only the projection of vector **E** in a direction in which it transmits the oscillations. As a result, when plate C is observed in monochromatic light through analyzer N_2, the plate will appear to be illuminated nonuniformly. In white light, the surface of plate C will appear to be coloured.

PHENOMENA OCCURRING IN CONVERGENT RAYS. Figure 250 presents the path of the rays in the plane of the figure. The crystal plate is cut perpendicular to the optical axis. It can be seen from the figure that the pattern of rays

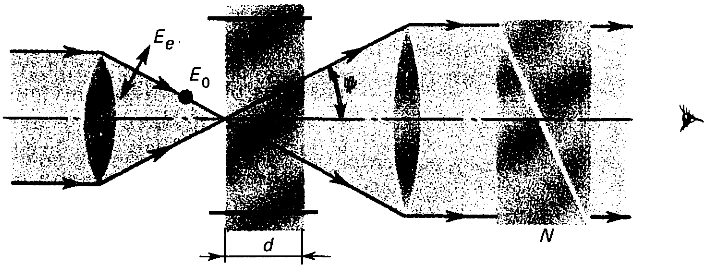


Fig. 250. To the analysis of phenomena taking place in convergent beams.

passing through the crystal plate is axially symmetric if the incident beam of light is nonpolarized.

The direction of oscillations of the electric vector of the extraordinary ray lies in the plane of the figure and is shown by arrows. For the ordinary ray, the electric vector oscillates in a direction perpendicular to the plane of the figure shown by dots. The path length of the ray in the plate is $h = d/\cos\psi$, where ψ is the angle between the ray and the axis. Hence the phase diference between the rays in the given direction is

$$\Delta = (d/\cos\psi)(n_e - n_o). \tag{43.12}$$

The phase difference between the rays emerging from the plate is given by

$$\delta = \frac{2\pi}{\lambda}\frac{d}{\cos\psi}(n_e - n_o), \tag{43.13}$$

and is different for different angles. Moreover, formula (43.13) shows that the pattern of rays is axially symmetric. Hence elliptically polarized rays with different orientations of the ellipses propagate from the plate at different angles to the axis. If the radiation is monochromatic and Nicol prism N_2 is placed in the path of the rays emerging from the plate, alternating dark and bright rings corresponding to the projections of the electric vector onto the principal direction of Nicol prism N_2 are observed. If the prism is turned through an angle of $\pi/2$, the bright rings are transformed into dark ones, and vice versa. If white light is used for illumination, a set of concentric coloured circles is observed.

If a Nicol prism N_1 is placed in the path of the incident rays, the pattern of the rays changes. The concentric circles are intersected by a cross which is bright if the principal directions of Nicol prisms N_1 and N_2 are parallel, and dark if these two directions are perpendicular. This is explained by the fact that the incident beam in the plane of incidence coinciding with the principal plane of Nicol N_1 contains only extraordinary rays, and hence there is no interference with ordinary rays. Hence the light leaving the plate is linearly polarized in it. If the principal direction of Nicol N_2 lies in this plane, the rays will pass this plane without attenuation, and hence a bright fringe is observed. If Nicol N_2 is at right angles to N_1, we observe a dark spot instead of a bright one.

In the same way, we can analyze rays in a plane perpendicular to the plane of the figure. For the same initial polarization, rays in this plane are only ordinary ones. A linearly polarized ray leaves the plate and passes through Nicol N_2 since the direction of oscillations of its vector **E** is the same as for an extraordinary ray in the perpendicular plane. This leads to the creation of a bright fringe which is perpendicular to the bright fringe formed by the extraordinary ray. These two bright fringes form together a cross which intersects the coloured concentric rings.

SEC. 44. ROTATION OF THE POLARIZATION PLANE

A phenomenological description of natural and artificial optical activity of substances is given.

ROTATION OF POLARIZATION PLANE IN CRYSTALLINE BODIES. When linearly polarized ray passes along the optical axis of a quartz plate (Fig. 251), the polarization plane turns (Arago, 1811). There is no splitting of a ray upon normal incidence on a plate cut perpendicular to the optical axis. The angle through which the polarization plane turns is determined by the mutual orientation of the axes of Nicols N_1 (polarizer) and N_2 (analyzer). It has been established experimentally that the angle of rotation depends on the path length d in the crystal plate, as well as on the wavelength, i. e. a rotatory dispersion takes place:

$$\varphi = \alpha d, \tag{44.1}$$

where α is the rotatory power. This quantity is expressed in radians per metre (in SI) or in degrees per millimetre (°/mm). For quartz, $\alpha \simeq 15°/\text{mm}$ in the red spectral region and $\alpha \simeq 27°/\text{mm}$ in the green region. For the violet region of the spectrum, the value of α is approximately 51°/mm. These results show that quartz has a sufficiently high rotatory power.

The direction of rotation of the polarization plane varies upon a reversal of the direction of propagation of light and can be determined with the help of the screw rule.

It has been established experimentally that there are two modifications of quartz crystals, viz. dextrorotatory and levorotatory. They are characterized by different directions of

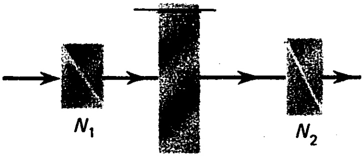


Fig. 251. Rotation of the polarization plane upon the passage of light along the optical axis through a quartz plate.

rotation of the polarization plane, which are determined by the right- or left-hand screw rule when light propagates along the optical axis. Conventionally, the direction of rotation of the polarization plane is determined with respect to the observer towards whom the ray of light is directed. Hence the dextrorotatory modification of quartz (Fig. 252*a*) causes an anticlockwise rotation of the polarization plane, while the levorotatory quartz (Fig. 252*b*) causes a clockwise rotation.

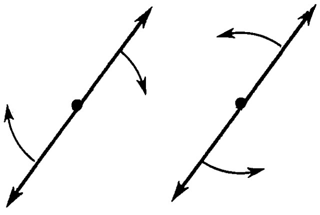


Fig. 252. Rotation of the polarization plane in dextrorotatory (*a*) and levorotatory (*b*) crystal.

It has been established that if a crystal rotates the polarization plane, it always contains both modifications, and

$$|\alpha_{(+)}| = |\alpha_{(-)}|, \tag{44.2}$$

where $\alpha_{(+)}$ and $\alpha_{(-)}$ are the rotatory powers of the dextrorotatory and levorotatory modifications of the crystal. (Henceforth, we shall call these modifications the right-handed and left-handed crystal respectively.)

ROTATION OF POLARIZATION PLANE IN AMORPHOUS SUBSTANCES. In many amorphous substances, also including liquids (for example, turpentine oil), the polarization plane is rotated. In some cases, the extent of rotation is quite significant (for example, $\alpha \simeq 1.64°/\text{mm}$ in nicotine for the yellow spectral region).

Amorphous substances also have two modifications, viz. left-handed and right-handed. If a substance causing a rotation of polarization plane is in solution, the angle of rotation of the polarization plane when the light passes through the solution is determined by the Biot law (1832):

$$\varphi = \alpha q l, \tag{44.3}$$

where q is the concentration of the solution, and l is the path length in the solution. The rotatory power α depends on the wavelength and the solution temperature. An experimental investigation of this phenomenon showed that α has a weak dependence on temperature, while its dependence on wavelength is roughly given by $\alpha \propto 1/\lambda^2$.

The rotation of the polarization plane in amorphous substances is a molecular phenomenon. A molecule can be assigned an appropriate rotatory power which is independent of the state of aggregation of matter. The crystal structure of a substance can also cause rotation. Materials capable of rotating the polarization plane are called optically active substances.

!

The rotation of polarization plane is due to the different velocities of right and left circularly polarized waves.

PHENOMENOLOGICAL THEORY OF THE ROTATION OF POLARIZATION PLANE. A linearly polarized wave can be treated as a superposition of two circularly polarized waves with opposite directions of rotation of the electric vector [see (5.10), (5.11) and Fig. 19]. The theory is

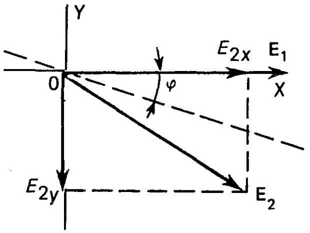


Fig. 253. To the calculation of the angle of rotation of the polarization plane.

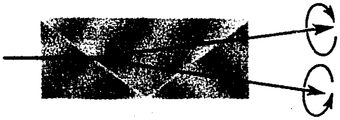


Fig. 254. Experiment proving the existence of circularly polarized waves in optically active substances.

based on the assumption that circularly polarized waves having different directions of rotation of the electric vector have different velocities (Fresnel, 1817). If a ray described by formulas (5.10) and (5.11) enters a crystal, a constant phase difference is created in the oscillations of vectors E_1 and E_2, which results in a rotation of the polarization plane.

Assuming that the waves (see Fig. 19) propagate in the positive direction of the Z-axis (towards us), and denoting by n_1 and n_2 the refractive indices of the circularly polarized waves with vectors E_1 and E_2 respectively, we can write, on the basis of (5.10) and (5.11), the following expressions for the waves leaving the crystal:

$$E_{1x} = E_0 \cos(\omega t - kn_1 d), \quad E_{1y} = E_0 \sin(\omega t - kn_1 d), \tag{44.4}$$

$$E_{2x} = E_0 \cos(\omega t - kn_2 d), \quad E_{2y} = -E_0 \sin(\omega t - kn_2 d), \tag{44.5}$$

where d is the distance traversed by the wave in the medium responsible for the rotation of the polarization plane. Obviously, vectors E_1 and E_2 are now symmetric not about the line $y = 0$ as in Fig. 19, but about a different line which determines the rotation of the plane of the oscillations of the resultant vector $\mathrm{E}_1 + \mathrm{E}_2$. To find the position of this plane, we determine the position of vectors E_1 and E_2 at a certain instant of time, say, $t = kn_1 d/\omega$. Then [see (44.4) and (44.5)]

$$E_{1x} = E_0, \quad E_{1y} = 0, \tag{44.6}$$

$$\begin{aligned} E_{2x} &= E_0 \cos[kd(n_1 - n_2)], \\ E_{2y} &= -E_0 \sin[kd(n_1 - n_2)]. \end{aligned} \tag{44.7}$$

The position of these vectors is shown in Fig. 253 for the case $(n_1 - n_2) > 0$. If $(n_1 - n_2) < 0$, then vector E_2 is turned in the other direction relative to vector E_1. The line about which vectors E_1 and E_2 are symmetric is the bisector of the angle between these vectors, denoted by 2φ in Fig. 253. Consequently, the polarization plane rotates through an angle

$$\varphi = k(n_1 - n_2)\, d/2. \tag{44.8}$$

Since the direction of rotation of vector E in a circularly polarized wave is usually characterized from the side of the observer towards which the ray is directed, n_1 is the refractive index of the wave with left-circular polarization, and n_2 is the corresponding quantity for right-circular polarization of the wave. Hence for a dextrorotatory substance $n_1 > n_2$, while for a levorotatory material, $n_1 < n_2$.

The validity of the theory is confirmed in experiments on the refraction of circularly polarized waves (Fig. 254). On account of the difference between the ratios of the velocities of right- and left-circularly polarized waves in dextrorotatory and

?

What are mirror isomers? Give examples of spatial structures of molecules which are mirror isomers. Which dextrorotatory and levorotatory mirror isomers, if any, exist in nature?

levorotatory media, birefringence takes place at their interface. Two circularly polarized waves, which can be separated in space (Fig. 254), are formed by an incident linearly polarized wave. It has been experimentally confirmed that these waves are indeed circularly polarized and the refraction at the interface of the two media takes place in accordance with Huygens' construction for velocities, which is obtained from an analysis of the rotation of the polarization plane. Thus the phenomenological theory of the rotation of polarization plane was confirmed experimentally.

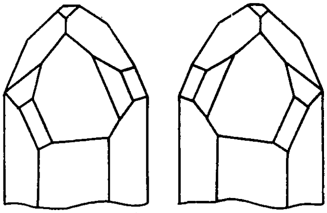


Fig. 255. Right-hand and left-hand quartz crystals.

OPTICAL ISOMERISM. The right-hand and left-hand crystals have identical chemical composition but different shapes such that they are mirror images of each other. Figure 255 shows the edges of a left- and right-hand quartz crystals. Similar difference is also observed between the molecules of amorphous optically active materials. Figure 256 presents two molecules in which atoms A and C are different. Obviously, these molecules are sort of mirror images of each other. Such molecules are called optical isomers. No spatial displacement can make an object coincide with its mirror image. Hence the right- and left-hand crystals or molecules exist in nature independently. In particular, the molecules of amino acids constituting living organisms on the Earth are optically active, and all of them are levorotatory molecules. Since the structure of molecules, especially organic compounds, is quite complicated, it is possible to form many different pairs of optical isomers.

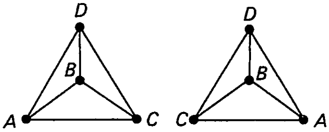


Fig. 256. Molecules with mirror symmetry.

ROTATION OF POLARIZATION PLANE IN A MAGNETIC FIELD. In a magnetic field, optically inactive materials become optically active and rotate the polarization plane of light propagating in a material along the magnetic field lines. This effect was discovered in 1846 by M. Faraday (1791-1867) and is called the Faraday effect. This effect describes the influence of a magnetic field on matter, making a substance optically active.

The diagram for observing the rotation of the polarization plane by optically inactive substances in a magnetic field $H_{\|}$ is shown in Fig. 257. The angle of rotation φ of the polarization plane is defined by the relation

$$\varphi = VH_{\|}l, \tag{44.9}$$

where V is the **Verdet constant** characterizing the properties of a substance and depending on the light frequency and temperature, and $H_{\|}$ is the projection of the magnetic field vector on the direction of light propagation.

Formula (44.9) describes the angle of rotation of the polarization plane in para- and diamagnetic materials.

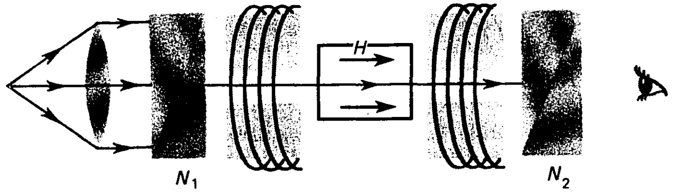


Fig. 257. Diagram for observing the rotation of polarization plane by optically inactive substances in a magnetic field.

The rotation angles φ are not large. For magnetic field strengths of the order of 1 MA/m and $l \simeq 0.1$ m, the rotation angle for most solids is about 1-2°, while for gases this value is even smaller.

In ferromagnetic materials, the rotation of the polarization plane is expressed in terms of magnetization J instead of the magnetic field strength. In this case, quantity H in formula (44.9) is replaced by J, and the Verdet constant by Kundt's constant, whose value is given in the tables.

The direction of rotation of the polarization plane is conventionally defined relative to an observer looking in the direction of the magnetic field vector (magnetization vector in the case of ferromagnetics).

Most materials are dextrorotatory and are called positive materials. These include all diamagnetics. There also exist negative materials and these always contain paramagnetic atoms.

The direction of rotation of the polarization plane in a magnetic field depends only on the direction of light propagation. In this respect, the rotation of the polarization plane in a magnetic field differs from the natural optical activity since it is caused by the action of the magnetic field on a substance.

The Verdet constant is determined from the Biot law:

$$V = A/\lambda^2 + B/\lambda^4, \qquad (44.10)$$

where A and B have a weak temperature dependence.

SEC. 45. ARTIFICIAL ANISOTROPY

Different methods of creating artificial anisotropy are described.

ANISOTROPY DUE TO DEFORMATION. An optically isotropic body subjected to compression or extension acquires the properties of a uniaxial crystal (Fig. 258) whose optical axis is collinear with the direction of the deforming forces. The following connection between the refractive indices of extraordinary and ordinary rays has been established experi-

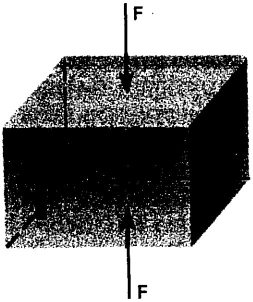


Fig. 258. Direction of the optical axis for compressive and extensive deformation.

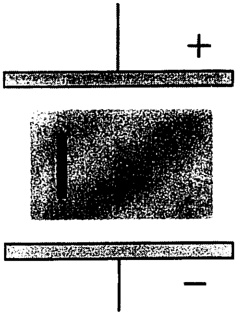


Fig. 259. Direction of the optical axis in a material in an electric field.

mentally for the direction perpendicular to the optical axis:

$$n_e - n_o = b\sigma . \tag{45.1}$$

Here $\sigma = F/S$ is the stress responsible for the deformation and b is a constant characterizing the properties of the material. The difference $(n_e - n_o)$ can assume positive or negative values; besides, it depends on the wavelength of light.

When a transparent object is viewed in the gap between crossed Nicol prisms, the deformed body appears to be coloured, and the colouring depends on the deformation. The deformation distribution in the body can be determined from the colour distribution in it.

ANISOTROPY CAUSED BY AN ELECTRIC FIELD. In an electric field (Fig. 259), an optically isotropic body acquires the properties of a uniaxial crystal with its optical axis collinear with the electric field vector (Kerr effect, 1875). The following relation between the ordinary and extraordinary waves has been established experimentally for the light propagating perpendicular to the optical axis:

$$n_e - n_o = K\lambda E^2 , \tag{45.2}$$

where K is the Kerr constant. For liquids, this constant is equal to a few picometres per square volt. Thus, for liquid oxygen at 90 K and liquid nitrogen at 71.4 K, the Kerr constant for a wavelength of 546 nm is equal to 10.1 and 4.02 pm/V^2 respectively. For gases, the value of this constant is significantly smaller. For example, it is equal to 0.45×10^{-15} and 0.3×10^{-15} m/V^2 for oxygen and nitrogen under normal conditions. This is due to the fact that the Kerr effect depends on the properties of molecules and is therefore enhanced as the number density of molecules increases. It can be seen from (45.2) that the optical properties of the material remain unchanged upon a reversal of the electric field. In other words, it really behaves like a uniaxial crystal.

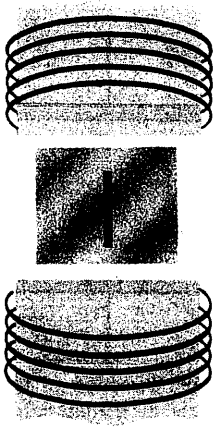


Fig. 260. Direction of the optical axis in a material in a magnetic field.

After traversing a distance l, the optical paths of ordinary and extraordinary rays differ by $\Delta = K\lambda/E^2$, while the phase difference between the waves is

$$\varphi = k\Delta = 2\pi K l E^2 . \tag{45.3}$$

For most liquids, $K > 0$. The Kerr constant also depends on the wavelength (i.e. dispersion effect takes place), and decreases with increasing temperature.

To explain the Kerr effect, we must take into account two physical factors. In an electric field, nonpolar molecules acquire a dipole moment in the direction of the field, while the molecule itself is oriented in such a way that its dipole moment coincides with the direction of maximum polarization of the molecule. Consequently, the maximum refractive index corresponds to the wave whose electric vector oscillates collinearly with the external electric field, i.e. to the extraordinary wave:

$$n_e > n_o , \quad K > 0 . \tag{45.4}$$

In an external electric field, polar molecules are oriented with their permanent dipole moments predominantly in the direction of the field. The direction of maximum polarization of a molecule generally forms a certain angle with the electric field in this case. Depending on this angle, various relations between n_e and n_o, or various values of K (including zero) are possible. The Kerr effect has a very small time lag (the optical anisotropy lags behind the variation in the electric field by about 10^{-10} s). Hence it is used in the construction of high-speed modulators of light called **Kerr cells**.

The size of a Kerr cell is determined in such a way that its optical length is equal to half the wavelength for a certain value of the electric field strength. If the cell is placed between two crossed Nicol prisms whose principal planes are included at an angle of 45° to the optical axis appearing in the Kerr cell in an electric field, the cell does not transmit light in the absence of an electric field. As the field is switched on, light begins to pass through the cell and at a certain value of the electric field the cell acts like a half-wave plate, and the intensity of the light passing through the system attains its maximum value.

ANISOTROPY CAUSED BY A MAGNETIC FIELD (Cotton-Mouton effect, 1910). In a magnetic field, an optically isotropic material acquires the properties of a uniaxial crystal whose axis is collinear with the magnetic induction vector (Fig. 260). The dependence of the difference $(n_e - n_o)$ on B when light propagates in a direction perpendicular to the magnetic induction is given by the relation

$$n_e - n_o = C\lambda B^2, \tag{45.5}$$

where C is a constant characterizing the properties of the material. The magnitude of this effect is quite small, and the essence of the effect is the same as for the Kerr effect. It must be observed that this effect is basically different from the Faraday effect (44.9) in that it depends on the square of the magnetic induction as against the linear dependence in the case of the Faraday effect.

The constant C is determined experimentally. For nitrobenzene, for example, $C = 2.25 \times 10^{-2}\ \mathrm{m}^{-1}\mathrm{T}^{-2}$. When light traverses a distance of 1 m in a magnetic field with induction 1 T, the phase difference is just 0.14 rad.

POCKELS EFFECT. The application of an external electric field to some crystals leads to birefringence. This effect does not exist in the absence of an electric field, and the difference in the refractive indices of ordinary and extraordinary rays is directly proportional to the electric field strength. This phenomenon is called the Pockels effect.

The optical axis of the crystal is oriented parallel to the ray of light, and the external electric field is also collinear with this direction. This can be ensured by either using transparent electrodes or drilling small holes in the centres of the electrodes. The application of an external electric field leads to the appearance of a second optical axis in a plane perpendicular to the first optical axis. Consequently, the induced optical axis is perpendicular to the direction of light propagation, and the birefringence of light takes place relative to this axis. For this axis, we have

$$n_e - n_o = aE, \tag{45.6}$$

where a is a constant. The orientation of the induced axis in a plane perpendicular to the direction of the ray of light depends on the orientation of the crystal. The voltage which must be applied to the electrodes in the Pockels effect is about an order of magnitude lower than for producing the same extent of birefringence in the Kerr effect (for the same distance between the electrodes). This is a significant advantage of the Pockels effect from a practical point of view. For example, the value of *a* for a potassium dihydrophosphate (KH_2PO_4) crystal is about 3.6×10^{-11} m/V, while for lithium niobate its value is 3.7×10^{-10} m/V. The Pockels effect is as inertialess as the Kerr effect and is used for creating high-speed modulators of light. Such a device is called the Pockels cell.

The Pockels effect arises not only when a field is applied along the optical axis of the crystal, but also when the field is applied in a perpendicular direction. In the former case, the effect is called the longitudinal Pockels effect, while in the latter case it is called the transverse Pockels effect. In both cases, however, the ray of light must propagate in the direction of the optical axis of the crystal in the absence of an applied electric field.

For the fabrication of the Pockels cells, the transverse effect has a distinct advantage over the longitudinal effect. In the first place, the electrodes are parallel to the beam of light and the distance between them can be made quite small while the length along the ray can be made quite large. Hence a half-wave cell can be made to operate for a comparatively small potential difference between the electrodes. In the case of longitudinal effect, the phase difference between ordinary and extraordinary waves is independent of the cell length for a fixed potential difference, since an increase in length entails a decrease in the electric field strength for a fixed phase difference. Hence the phase difference can be increased only by increasing the potential difference between the electrodes. Secondly, it is technically easier to fabricate a cell operating on the basis of the transverse effect than on the longitudinal effect. Hence the low-voltage Pockels cells are usually based on the transverse effect. However, it is expedient to use the longitudinal effect for creating high-speed cells, since in this case the electrodes have a smaller size and a lower capacitance, and this ensures a rapid variation of potentials.

The Pockels cells are applied for the same fields as the Kerr cells and have replaced them in many cases.

PROBLEMS

8.1. A crystal plate is placed between two Nicol prisms whose principal planes form angles α and β with the optical axis of the crystal. The optical axis is parallel to the plane of the crystal on which light is incident. Find the intensity of light emerging from the crystal. The thickness of the plate is *d*.

8.2. A quartz plate of thickness 0.014 mm is cut parallel to the optical axis. For $\lambda = 0.5$ μm, the refractive indices are $n_e = 1.5533$ and $n_o = 1.5442$. Linearly polarized light is incident on the plate and the plane of oscillations of the electric vector forms an angle of 25° with the optical axis. Find the nature of polarization of the light emerging from the plate.

8.3 A birefringent prism (Fig. 242) consists of a glass prism with $n = 1.66$ and a prism of Iceland spar with $n_o = 1.66$ and $n_e = 1.49$. The angle α is equal to 30°. Find the angle between the rays emerging from the prism if the ray enters the prism as shown in Fig. 242.

ANSWERS

8.1. $I = I_0 [\cos^2(\alpha - \beta) - \sin 2\alpha \sin 2\beta \sin^2(\delta/2)]$; $\delta = k(n_e - n_o)d$.
8.2 Circular polarization.
8.3. 5°45′.

Chapter **9**

Scattering of Light

Basic idea:
Under the effect of incident radiation, atoms and molecules of a substance, as well as groups of atoms and molecules become sources of secondary radiation.

SEC. 46. SCATTERING PROCESSES

The essence and classification of various processes of scattering of light are described.

NATURE OF SCATTERING. A medium contains molecules or atoms of the basic material, as well as particles of foreign matter (dust, water drops, etc.). The size of the basic molecules is about 0.1 nm, while foreign particles consist of aggregates of molecules and are thousands or tens of thousands times larger in size.

In the process of scattering, a molecule or a particle borrows energy from an electromagnetic wave propagating in the medium, and then emits this energy within a solid angle whose vertex is the molecule or the particle itself. In this sense, the scattering of light by a molecule or a particle formed by a very large number of molecules proceeds identically, and the only difference lies in the mechanism of re-emission of energy.

If the medium is treated as a continuum, optical inhomogeneities of the medium are responsible for scattering. In this case, the medium is phenomenologically characterized by a varying refractive index, and the "size" of the regions over which scattering takes place is determined by the distances over which the refractive index changes significantly. In the physical sense, scattering of light is just the diffraction of a wave at the inhomogeneities of the medium.

!

The process of scattering is reduced to the generation of secondary waves by molecules or particles under the effect of incident radiation. For a continuous medium, scattering is reduced to the diffraction of waves at the inhomogeneities of the medium.

TYPES OF SCATTERING. In the first place, the nature of scattering depends on the ratio of the wavelength to the particle size. If the particle size is less than about 1/15 of the wavelength, the scattering is called the Rayleigh scattering after Rayleigh (1842-1919) who first studied this type of scattering. For larger particle size, the Mie scattering takes place. Although the original theory of this phenomenon was developed by Mie (1868-1957) for spherical particles only, the term “Mie scattering” is also applicable for particles of irregular shape. For small particles, the Mie scattering theory leads to results obtained from the Rayleigh scattering theory. An important particular case of optical inhomogeneity is the inhomogeneity in the optical properties of a medium through which an acoustic wave is propagating. This results in a harmonic distribution of the optical inhomogeneity of the medium in space, and a harmonic variation of the optical properties in time. The harmonic inhomogeneity of optical properties in space leads to diffraction of light by waves (see Sec. 33), while a harmonic variation of the optical properties in time leads to a variation in the frequency of the diffracted light at each point in the medium. This change in the frequency of light diffracted by an acoustic wave is called the Brillouin scattering. It was discovered by Brillouin (1854-1948) and independently by Mandelshtam (1879-1944).

The quantum-mechanical properties of molecules are manifested in the Raman scattering of light, characterized by a change in the frequency of scattered light in comparison to the frequency of incident light. In view of the specifically quantum nature of this phenomenon, it is treated as a special type of scattering.

MULTIPLE SCATTERING. The radiation scattered by a particle may in turn be scattered by another particle, and so on. In this case, we speak of multiple scattering. Multiple scattering is realized in accordance with the laws of simple scattering at each successive act. The final result is obtained by successively adding the results of simple scatterings taking into account the statistical nature of their occurrence.

SEC. 47. RAYLEIGH SCATTERING AND MIE SCATTERING

The basic principles of the Rayleigh- and Mie scatterings are described.

MODEL OF AN ELEMENTARY SCATTERER. Under the influence of the electric field of an electromagnetic wave electrons oscillate at the frequency of the electromagnetic wave. If the wave propagates in the positive direction of the X-axis (Fig. 261), and the electric vector oscillates in the plane $Y = 0$, the equation of motion of the electron has the following form in accordance with (9.13):

$$m\ddot{z} + m\omega_0^2 z = eE_0 \cos \omega t, \tag{47.1}$$

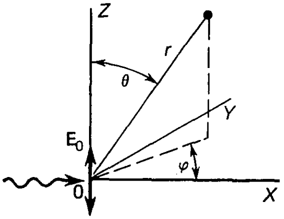


Fig. 261. Elementary scatterer of light.

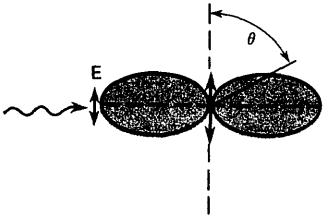


Fig. 262. Angular distribution of the intensity of scattered radiation from a polarized wave.

where $E_0 \cos \omega t$ describes the oscillations of the electric field strength of the wave which is collinear with the Z-axis in the plane $Y=0$; m and e are the mass and charge of the electron; and ω_0 is the natural frequency of oscillations of the electron, determined by the elastic force confining the electron to the equilibrium position. In view of its small magnitude, the damping of electron oscillations is neglected.

From (47.1), we obtain the following expression for the deviation of the electron from its equilibrium position:

$$z = \frac{e}{m} \frac{E_0}{\omega_0^2 - \omega^2} \cos \omega t . \tag{47.2}$$

An oscillating electron itself is an emitter of radiation, its radiation being scattered. Thus, the model of an ideal classical scatterer of light is the elementary classical emitter, viz. an electric dipole, in the field of the electromagnetic wave.

The electron forms a part of an atom which is electrically neutral. Hence it can be assumed that in accordance with (47.2), the electron oscillates about the equilibrium position in which the positive charge $|e|$ is located. This charge can be assumed to be fixed since the mass of the proton (or nucleus) bearing this charge is much more than the electron mass. Consequently, (47.2) can be written as a formula for the variation of dipole moment with time:

$$p_z(t) = ez = \frac{e^2}{m} \frac{1}{\omega_0^2 - \omega^2} E_0 \cos \omega t . \tag{47.3}$$

In the spherical system of coordinates with the polar axis coinciding with the direction of the dipole, the field of the electromagnetic wave emitted by the dipole is given by the formulas (Fig. 262)

$$E_\theta = cB_\varphi = -\frac{1}{4\pi\varepsilon_0} \frac{\omega^2}{c^2} \frac{\sin \theta}{r} p(t - r/c),$$

$$E_r = E_\varphi = 0, \quad B_r = B_\varphi = 0, \tag{47.4}$$

where θ and φ are the polar and axial angles respectively, and r is the distance from the dipole to the point at which the field is determined (see Fig. 261). It should be emphasized that there is no radiation along the line of oscillations of the dipole.

The energy flux density in the direction characterized by the angles θ and φ is given, in accordance with (3.1), by

$$S(\theta, \varphi, r, t) = E_\theta H_\varphi = \frac{E_\theta B_\varphi}{\mu_0} = \frac{\omega^4 \sin^2\theta p^2}{16\pi^2 \varepsilon_0 c^2 r^2} . \tag{47.5}$$

Averaging S over a period and taking (47.3) into consideration, we obtain

$$\langle S\rangle_t = \frac{\pi^2 c\sin^2\theta}{2\varepsilon_0 r^2\lambda^4}\left[\frac{e^2}{m(\omega_0^2-\omega^2)}\right]^2 E_0^2, \tag{47.6}$$

where $\lambda = 2\pi c/\omega$.

The energy flux $\mathrm{d}P(\theta,\varphi)$ over a solid angle $\mathrm{d}\Omega = \mathrm{d}\sigma/r^2$ formed by an area element $\mathrm{d}\sigma$ of a sphere is

$$\mathrm{d}P(\theta,\varphi) = \langle S\rangle_t\mathrm{d}\sigma = \langle S\rangle_t r^2\mathrm{d}\Omega. \tag{47.7}$$

This leads to the following expression for the scattering intensity $I_1(\theta,\varphi)$, defined as the energy flux corresponding to the solid angle $\mathrm{d}\Omega$:

$$I_1(\theta,\varphi) = \frac{\mathrm{d}P}{\mathrm{d}\Omega} = \frac{\pi^2 c\sin^2\theta}{2\varepsilon_0\lambda^4}\left[\frac{e^2}{m(\omega_0^2-\omega^2)}\right]^2 E_0^2. \tag{47.8}$$

It should be observed that (47.8) expresses the energy density of the scattered flux from one elementary emitter.

With the help of (3.4), we can transform this formula into

$$I_1(\theta,\varphi) = \frac{\pi^2\sin^2\theta}{\lambda^4}\left[\frac{e^2}{m(\omega_0^2-\omega^2)}\right]^2\langle S_0\rangle, \tag{47.9}$$

where $\langle S_0\rangle$ is the mean energy flux density in the incident wave.

RAYLEIGH SCATTERING. If the size of a scatterer is much smaller than the wavelength of the radiation, all elementary dipoles emit coherently. Usually, by the Rayleigh scattering we mean the scattering by molecules of a medium, since the size of ordinary molecules (not macromolecules) is always much smaller than the wavelength of visible light. Elementary scatterers belonging to different molecules emit incoherent radiation, since the distance between molecules can be quite large on the one hand, and the density of the medium can fluctuate due to molecular motion on the other hand. Taking these circumstances into account, we conclude that the intensity of a wave scattered by a molecule increases as the square of the number N_0 of elementary scatterers contained in it. If we denote the number density of molecules by N, there will be N_0N elementary dipoles in a unit volume of the medium. From the course on electricity, we know the relation

$$\frac{(n^2-1)3\varepsilon_0}{N_0N(n^2+2)} = \frac{e^2}{m(\omega_0^2-\omega^2)}, \tag{47.10}$$

where n is the refractive index of the medium. Replacing the expression in the brackets on the right-hand side of (47.9) by its value (47.10), we obtain the following formula for the scattering intensity due to one molecule:

$$I(\theta, \varphi) = I_1(\theta, \varphi) N_0^2 = \frac{9\pi^2 \sin^2\theta}{N^2\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2 \langle S_0 \rangle . \tag{47.11}$$

It should be noted that the formulas obtained here are valid for the case when the natural frequency ω_0 of the electron oscillations is much higher than the oscillation frequencies of the visible and near ultra-violet regions of the spectrum. In most cases, this condition is satisfied. Further, while deriving formula (47.12) we assumed that all electrons in a molecule have the same natural oscillation frequency. This restriction can be easily lifted since it is of no fundamental importance, and its removal simply makes the formulas more cumbersome. We leave it to the reader to make these calculations by way of an exercise.

The total intensity of scattering in all directions due to a single particle is obtained from (47.11) by integrating over all scattering angles:

$$I_0 = \int_{4\pi} I(\theta, \varphi)\, \mathrm{d}\Omega = \frac{18\pi^3}{N^2\lambda^4}\left(\frac{n^2-1}{n^2+1}\right)^2 \langle S_0 \rangle \int_0^{\pi} \sin^3\theta \mathrm{d}\theta , \tag{47.12}$$

$$I_0 = \frac{24\pi^3}{N^2\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2 \langle S_0 \rangle . \tag{47.13}$$

Since scattering by different molecules is not coherent, the total scattering intensity per unit volume of a material is calculated by multiplying (47.13) by the number density N of the molecules of the substance. For gases with a moderate number density, the refractive index $n \simeq 1$, and hence we can put $n^2 + 2 \simeq 3$ and $(n^2 - 1) \approx 2(n-1)$ in formula (47.13). In this case, we obtain the following expression for the scattering intensity per unit volume:

$$I_V = I_0 N = \frac{32\pi^3 (n-1)^2}{3N\lambda^4} \langle S_0 \rangle . \tag{47.14}$$

RAYLEIGH LAW. It can be seen from (47.14) that the scattering intensity is inversely proportional to the fourth power of the wavelength. Such a dependence of scattering on wavelength is called the Rayleigh law.

Among other things, the Rayleigh law explains the blue colour of the sky and the reddish colour of the sun at sunrise and sunset. In accordance with the Rayleigh law, the colour of the light scattered at sunrise or sunset has a much strongly attenuated short-wave part of the spectrum (violet) as a result of scattering than the long-wave (red) part. Hence the intensity of the long-wave (red) part of the spectrum increases relatively

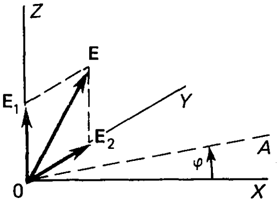


Fig. 263. To the calculation of scattering of nonpolarized radiation.

and is perceived by the eye as the crimson colour of the Sun. Relative variation of the intensity of various parts of the spectrum will be noticeable only in the case of a quite large scattering. Hence at the zenith, when the solar rays have to traverse a comparatively thinner layer of atmosphere and the scattering of light is insignificant, the Sun is not red in colour. However, even in this case the scattering and absorption considerably change the spectral composition of the radiation reaching the Earth's surface (see Sec. 1).

As we look at the sky in daytime, the scattered light falling on the eye contains predominantly the short-wave part of the spectrum, corresponding to the blue colour. Beyond the Earth's atmosphere, the sky appears to be black and only direct rays from the stars fall on the eye.

ANGULAR DISTRIBUTION AND POLARIZATION OF LIGHT IN RAYLEIGH SCATTERING. The angular distribution of scattering of polarized radiation from an individual molecule is described by formula (47.11). This distribution is axially symmetric with respect to the line passing through an elementary scatterer in the direction of the electric vector oscillations of the incident wave (Fig. 262). There is no scattering perpendicular to the direction of propagation of the incident wave, viz. along the line of oscillations of E. The maximum scattering is observed in the plane normal to the oscillations of the electric vector of the incident wave. The scattered radiation is polarized, i.e. the electric vector oscillates in a plane passing through the line along which the electron of an elementary scatterer oscillates. If the scattering from different molecules can be assumed to be incoherent, the total scattering intensity per unit volume is the product of expression (47.11) and the number density N of molecules. Hence the properties of radiation scattered by a single molecule are completely preserved for radiation scattered in the bulk of the medium.

To calculate the angular distribution of scattering of a nonpolarized wave, we can present the incident wave as a superposition of two waves with mutually perpendicular directions of polarization (see Sec. 4):

$$\mathrm{E} = \mathrm{E}_1 + \mathrm{E}_2, \tag{47.15}$$

where $\mathrm{E}_1 \perp \mathrm{E}_2$. Hence we obtain the following relation between the energy flux densities of these waves:

$$\langle S_0 \rangle = \langle S_{01} \rangle + \langle S_{02} \rangle. \tag{47.16}$$

Obviously, the scattering of nonpolarized light must be axially symmetric with respect to the direction of the ray of light. Hence it is sufficient to find the intensity of the scattered light

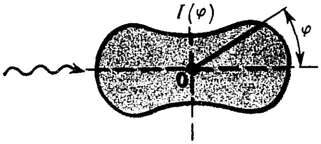


Fig. 264. Angular distribution of the intensity of scattering of nonpolarized light.

in the XY plane (Fig. 261) in a direction forming an angle φ with the X-axis (Fig. 263). We direct vectors E_1 and E_2 along the Z- and Y-axes respectively. This gives [see (47.11)]

$$I = \frac{9\pi^2}{N^2\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2 [\langle S_{01}\rangle \sin^2(\pi/2) + \langle S_{02}\rangle \cos^2\varphi], \tag{47.17}$$

since the contributions from waves with mutually perpendicular polarizations to the total intensity are independent of each other. Vector E of nonpolarized light has an equiprobable distribution over all angles in the YZ plane. Hence, averaging E^2 over angles, we obtain from (47.16)

$$\langle\langle S_0\rangle\rangle = \langle\langle S_{01}\rangle\rangle + \langle\langle S_{02}\rangle\rangle, \tag{47.18}$$

where the outer angle brackets indicate averaging over the directions of E. Obviously,

$$\langle\langle S_{02}\rangle\rangle = \langle\langle S_{01}\rangle\rangle = \langle\langle S_0\rangle\rangle/2. \tag{47.19}$$

Consequently, formula (47.17) assumes the form

$$I(\varphi) = \frac{9\pi^2}{2N^2\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2 (1+\cos^2\varphi)\langle S_0\rangle. \tag{47.20}$$

The dependence of I on φ is shown in Fig. 264. The scattering pattern is axially symmetric with respect to the direction of propagation of the incident wave. Forward and backward scatterings have identical intensities and are distributed symmetrically with respect to the scattering centre.

The term with $\langle S_{01}\rangle$ in (47.17) describes a scattered linearly polarized wave whose electric vector oscillates in the direction of the Z-axis, while the term with $\langle S_{02}\rangle$ describes a linearly polarized wave whose electric vector oscillates along the Y-axis. To make the scattering independent of the coordinate system, we choose for the plane of observation the plane passing through the incident ray and the point of observation. It can be stated that the term with $\langle S_{01}\rangle$ in (47.17) describes a scattered wave whose electric vector oscillates in a plane perpendicular to the plane of observation, while the term with $\langle S_{02}\rangle$ describes a wave whose electric vector oscillates in the plane of observation. In the last parentheses on the right-hand side of (47.20), the unity describes the scattering of a wave whose electric vector is perpendicular to the plane of observation, while $\cos^2\varphi$ stands for the scattering of a wave whose electric vector is parallel to the observation plane. Thus, a partially polarized radiation is obtained as a result of scattering of nonpolarized light. The degree of polarization of the scattered radiation depends on the angle φ and is defined by the relation

$$P = \frac{I_{\perp}(\varphi) - I_{\parallel}(\varphi)}{I_{\perp}(\varphi) + I_{\parallel}(\varphi)} = \frac{\sin^2\varphi}{1 + \cos^2\varphi}, \tag{47.21}$$

which means that both polarization components have the same intensity only when $\varphi = 0$ or $\varphi = \pi$. For all other values of φ, a scattering for which the electric vector is perpendicular to the observation plane is more intense.

It follows from (47.21) that light is totally polarized ($P = 1$) in the direction perpendicular to the incident ray. In particular, if the Rayleigh scattering alone takes place in the Earth's atmosphere, the sunlight scattered in it and propagating in the plane perpendicular to the Earth-Sun direction must be totally polarized. In actual practice, total polarization is never observed since the Rayleigh scattering of sunlight is always accompanied by the Rayleigh scattering of light diffusively reflected from the Earth's surface, as well as by the Mie scattering by aerosol particles in the air. A weak optical anisotropy of molecules constituting the air is also partly responsible for the depolarization of the observed light.

ATTENUATION OF LIGHT INTENSITY. The energy flux density of the light propagating in a medium decreases as a result of scattering. If a light beam has a cross section σ, it encounters $N\sigma dx$ scatterers over a path length dx. Each of these scatterers knocks out the power (47.13) from the beam. As a result, the energy flux in the beam decreases:

$$d\langle S_0\rangle\sigma = -I_0 N\sigma dx. \tag{47.22}$$

Combining (47.13) and (47.22) we can write

$$d\langle S_0\rangle = -\gamma\langle S_0\rangle dx, \tag{47.23}$$

where the scattering coefficient is given by

$$\gamma = \frac{24\pi^3}{N\lambda^4}\left(\frac{n^2 - 1}{n^2 + 1}\right)^2. \tag{47.24}$$

Hence the law of attenuation of the energy flux density in a light beam undergoing scattering is written in the form

$$\langle S_0(x)\rangle = \langle S_0(0)\rangle e^{-\gamma x}. \tag{47.25}$$

MIE SCATTERING. The Rayleigh theory correctly describes scattering by molecules as well as by tiny spherical particles whose radius is smaller than about 0.03λ. As the particle size increases, departures from the predictions of the Rayleigh theory become noticeable, and we have to resort to the Mie theory. The Mie scattering theory takes into account the particle size and expresses scattering in the form of a series in which the small parameter is given by

$$\alpha = ka = 2\pi a/\lambda, \qquad (47.26)$$

a being the radius of the spherical particle. Strictly speaking, the Mie scattering theory is applicable only to spherical particles. However, the term "Mie scattering" is also applied to particles of other shapes. It can be seen from (47.26) that it is the ratio of the particle size and the wavelength that is important and not the absolute size of the particle. The laws of scattering of the ultraviolet part of the spectrum by particles of a definite radius are the same as for the scattering of the red part of the spectrum by particles that are twice as large. For $\alpha < 0.2$, the Rayleigh scattering becomes predominant. In other words, the Rayleigh scattering is a limiting case of the Mie scattering.

ANGULAR INTENSITY DISTRIBUTION AND POLARIZATION OF RADIATION IN MIE SCATTERING. Large particles contain a large number of molecules. An electromagnetic wave induces oscillations of the elementary dipoles of each molecule. These oscillating dipoles become the emitters of secondary waves that comprise the scattered radiation. In this respect, the mechanism of the Mie scattering is analogous to the Rayleigh scattering mechanism. The differences between the two types of scattering are due only to the following circumstances.

1. In the Rayleigh scattering, all elementary scatterers are in the field of the same wave and emit coherent radiation. For the Mie scattering, we must take into consideration the re-emission of the primary wave by the elementary scatterers which are generally in different electromagnetic fields. In other words, we must consider the fact that the refractive index is not equal to unity in the volume of a particle.

2. In the Rayleigh scattering, interference takes place between the radiations emitted by the elementary scatterers of the same particle (molecule) with the same phase difference, irrespective of the direction. In the Mie scattering, we must take into account the difference in the phases of the radiation emitted by elementary scatterers, as well as the phase difference introduced into the observed radiation by the finite distance between elementary scatterers. The latter situation is responsible for a significant dependence of the radiation intensity distribution on direction, which represents the dependence of the conditions for interference between radiations from elementary scatterers on their mutual arrangement relative to the point of observation.

!

Scattering is determined by the size of particles at which the light is scattered. When the linear dimensions of particles constitute about 1/10 of the wavelength, the Rayleigh scattering is observed, while for larger particles the Mie scattering takes place. The intensity of the Rayleigh scattering is inversely proportional to the fourth power of the wavelength. The intensity of the Mie scattering depends weakly on the wavelength and is completely independent of the wavelength for particles whose linear dimensions are much larger than the wavelength.

?

Describe the angular distribution and polarization of light in Rayleigh scattering.
In which direction is the light totally linearly polarized in Rayleigh scattering? What is the physical reason behind this phenomenon?
Describe the angular distribution and the polarization of light in Mie scattering.
Why has the Mie scattering a very weak dependence on the wavelength of light if the particles at which scattering takes place are quite large?

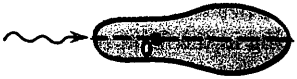


Fig. 265. Mie scattering for $a \approx \lambda/4$.

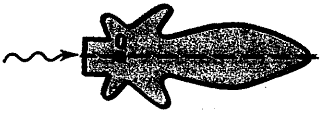


Fig. 266. Mie scattering for $a > \lambda$.

In the mathematical sense, the Mie scattering theory boils down to the solution of Maxwell's equations with boundary conditions on the surface of a spherical particle of an arbitrary radius, characterized by permittivity, permeability and conductivity. The solution is obtained in the form of series which contain complete information about the nature of scattering. On the whole a very cumbersome and complicated theory is obtained, but there is no need to describe it in this book. We shall confine ourselves to just a few important results.

As the particle size (a/λ, to be more precise) increases, an asymmetry in the forward and backward scattering is observed: forward scattering dominates (Fig. 265), but no sharp maxima or minima are observed (for $a \approx \lambda/4$). As the particle size increases further ($a > \lambda$), a predominantly forward scattering with a large number of secondary maxima whose distribution depends on the particle size (Fig. 266) takes place.

The scattered light is partially polarized even when the incident light is nonpolarized (as in the case of the Rayleigh scattering). The nature of polarization depends on the optical properties of the particles and the direction in which the scattered light is observed. If the incident light is polarized, the polarization of the scattered light depends on this polarization as well.

An important property of the Mie scattering is that it has a weak dependence on the wavelength for particles having a size considerably larger than the wavelength. This is significantly different from the Rayleigh scattering, and because of this the clouds are white and the sky is blue.

MANIFESTATIONS OF MIE SCATTERING. In view of the diversity of particles responsible for the Mie scattering, the latter is manifested in many different ways. The sky, which is blue at the zenith, gradually becomes grey towards the horizon. When the atmosphere is foggy, the sky acquires a whitish tinge. When an aeroplane flies very high, no sharp line of the horizon is usually observed, since it is covered by the atmospheric haze. All these phenomena are due to the Mie scattering by the aerosols in the atmospheric air. The low or nearly complete polarization of fog is the consequence of a strong Mie scattering by tiny water drops. The sharp attenuation of light from the Sun at dawn or dusk is also largely due to the Mie scattering.

SEC. 48. BRILLOUIN SCATTERING

Salient features of the Brillouin scattering are described.

BRILLOUIN COMPONENTS. Diffraction at an acoustic wave leads to the appearance of only two first-order maxima described by formula (33.64). The amplitude of the diffracted wave varies with the transmission coefficient and the refractive index of the medium, which vary due to a change in the density of the medium in the wave. For this reason the amplitude varies harmonically with the frequency Ω of the acoustic wave. Hence the electric field strength of an electromagnetic wave observed in the direction of the diffraction maxima is described by the formula

$$E(t) = A_0 \cos \Omega t \cos \omega t = (A_0/2)\left[\cos(\omega + \Omega)t + \cos(\omega - \Omega)t\right], \qquad (48.1)$$

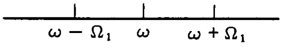


Fig. 267. Brillouin components.

where ω is the frequency of the incident light. Hence the scattered light must contain two satellite frequencies arranged symmetrically about the basic frequency (Fig. 267). The satellite line with frequency $\omega - \Omega$ is called the Stokes line, and the one with frequency $\omega + \Omega$ is called the anti-Stokes line. They are components of the Brillouin scattering.

Since the optical path length in a medium is n times the geometrical path length (n is the refractive index of the medium), formula (33.64) taking into account the variation in the velocity of light propagation in the medium can be presented by the equalities

$$nd \sin \varphi_{(+)} = \lambda, \quad nd \sin \varphi_{(-)} = -\lambda. \qquad (48.2)$$

The acoustic wave frequency can be expressed as follows:

$$\Omega = 2\pi v/d = 2\pi vn \sin \varphi/\lambda = 2\omega n (v/c) \sin(\varphi/2), \qquad (48.3)$$

where v is the velocity of sound waves in the medium, $\sin \varphi \approx 2 \sin(\varphi/2)$ in view of the fact that φ is small. Formula (48.3) is called the Brillouin formula.

UNDISPLACED COMPONENT. In most cases, besides the frequencies $\omega + \Omega$ and $\omega - \Omega$, the frequency ω is also observed in liquids. The appearance of this frequency can be formally explained if (48.1) contains $\alpha + \cos \Omega t$ instead of $\cos \Omega t$, where α is approximately constant. In other words, the presence of undisplaced frequency ω in the diffracted light is due to an optical characteristic of the medium which does not change according to a harmonic law but is practically constant. Such a constant component of optical inhomogeneity

?

How are the Brillouin components formed?
What physical factors are responsible for the undisplaced frequency in the scattering spectrum?
What are the main features of the Brillouin phenomenon in solids?

arises due to fluctuations in the medium, which level out over a period of time that is small in comparison with the period of the acoustic wave. As an example, we can mention the entropy fluctuations which level out on account of thermal conductivity.

BRILLOUIN SCATTERING IN SOLIDS. In amorphous solids, transverse as well as longitudinal waves can appear and propagate at different velocities. Each such wave generates two satellites in the scattered light. Hence the scattered radiation contains a total of five components including the undisplaced one. In crystalline solids, the number of components increases in accordance with the number of waves travelling at different velocities and in different directions, as well as the number of electromagnetic waves which can propagate in the crystal in a given direction.

Calculations show that a total of 24 displaced components appear in a crystal in the general case.

SEC. 49. RAMAN SCATTERING

Salient features of Raman scattering are described.

CLASSICAL INTERPRETATION. Suppose that the optical properties of a molecule vary according to a harmonic law, and consequently the amplitude of the light scattered by the molecule also changes in harmonic fashion. In analogy with (48.1), the observed electric field strength of the scattered light is given by

$$E(t) = A_0(1 + \alpha \cos \Omega t) \cos \omega t, \tag{49.1}$$

where Ω is the frequency characterizing the variation in the optical properties of the molecule and ω is the frequency of the light incident on the molecule. The coefficient α takes into account the efficiency of the amplitude modulation of the incident light by the molecule.

It can be seen from (49.1) that the scattered radiation contains waves with frequencies ω, $\omega + \Omega$ and $\omega - \Omega$. The existence of displaced frequencies in the radiation scattered by a molecule is due to the Raman scattering. This effect was discovered in 1928 by Raman and independently by Landsberg and Mandelshtam. Each spectral line of the primary radiation is accompanied in the scattered radiation by a whole set of satellites whose frequencies differ from the central frequency by

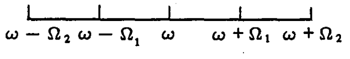


Fig. 268. Arrangement of satellite frequencies in Raman scattering.

amounts that are characteristic for a molecule (Fig. 268). It can be stated that a molecule has a set of natural frequencies Ω_1, Ω_2, ... of the oscillations of its optical properties. These frequencies are manifested in the scattering spectrum in accordance with formula (49.1). The frequencies of the scattered light are obtained as combinations of the incident light frequency and the natural frequencies of the molecular oscillations.

EXPERIMENTAL FACTS. It has been shown experimentally that the satellites accompany each spectral line of the incident radiation and that the frequencies Ω_1, Ω_2, ... are identical for all lines and characterize the properties of the molecules. The satellite lines are symmetric about the frequency of the incident radiation.

Satellites on the side of higher frequencies are called violet or anti-Stokes satellites, while those on the side of lower frequencies are called red or Stokes satellites. It has been established experimentally that the Stokes satellites nearest to the central frequency are much more intense than the anti-Stokes satellites.

However, this difference in intensities disappears as the temperature increases, since the intensity of the anti-Stokes lines increases much more rapidly with temperature.

The emission spectra of molecules are called band spectra since they are in the form of bands consisting of closely spaced lines.

This form of the spectra is due to the spreading of the electron line spectra of molecular radiation as a result of transitions between the vibrational and rotational energy levels of molecules. The energy gap between vibrational levels is much larger than between the rotational levels, and hence a band spectrum is formed as if in two stages. Vibrational spectral lines are formed at certain distances from the emission frequency as a result of electron transitions. Near each line of the vibrational spectrum, very closely spaced lines are formed due to rotational transitions. An analysis of the emission spectra of molecules as well as their Raman scattering spectra showed that the Raman frequencies Ω_1, Ω_2, ... always coincide with the difference in the frequencies of vibrational spectrum of the molecules. In other words, the Raman frequencies are equal to the natural oscillation frequencies of molecules. However, not all natural frequencies of the vibrating molecules correspond to the Raman frequencies in the Raman scattering spectrum.

Moreover, there is no simple relation between the intensity of the absorption line in a molecular spectrum and the corresponding line in the Raman spectrum.

?

What physical factors are responsible for band spectra of molecules?
Why are the Stokes satellites nearest to the central frequency much more intense than anti-Stokes satellites?
Why does this difference decrease with increasing temperature?
Why cannot the Raman frequency in the Raman spectrum be associated with all natural frequencies of molecular vibrations?

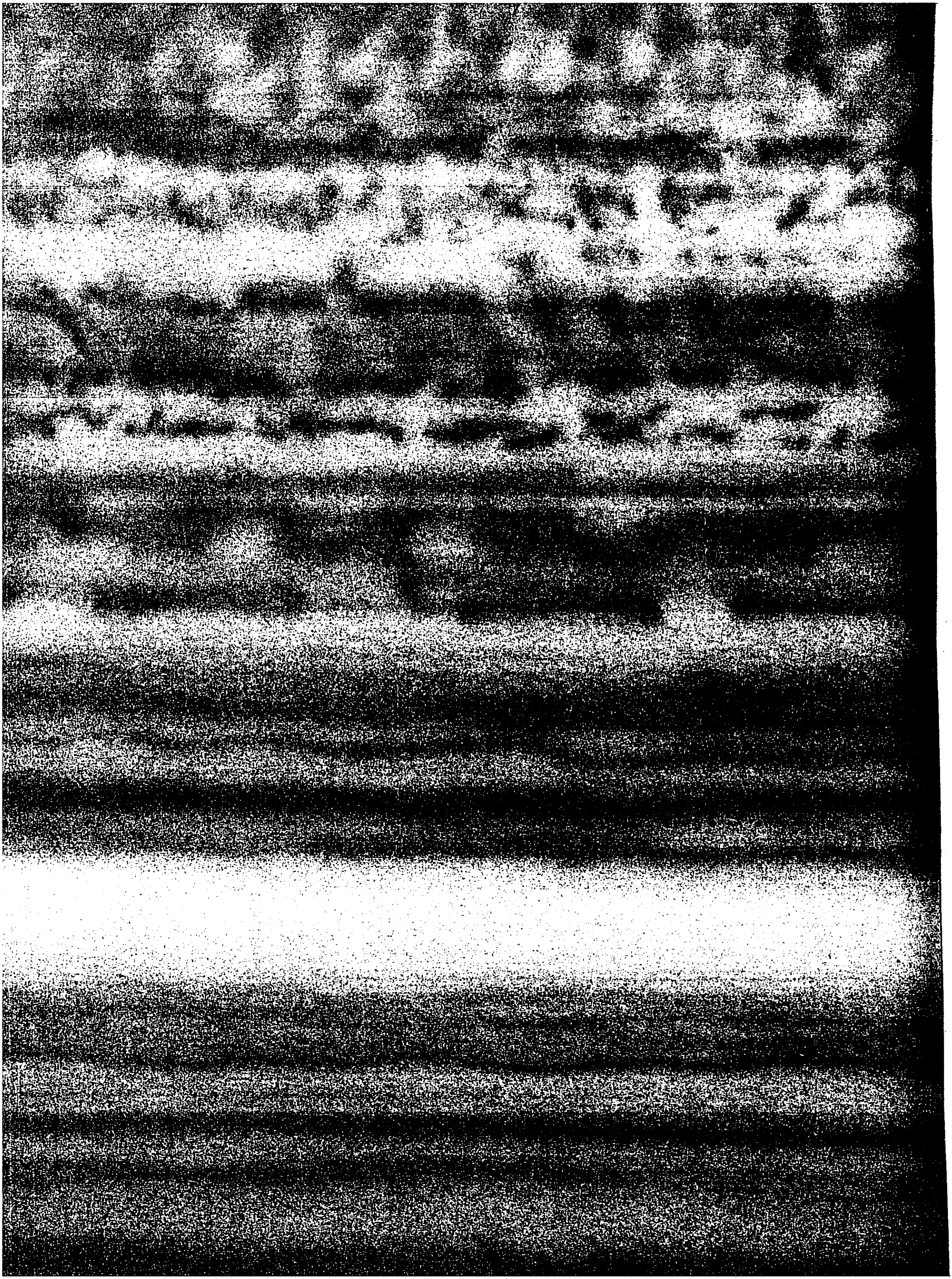
The classical interpretation (49.1) of Raman scattering helps in understanding the meaning of the Raman frequencies, but is not in a position to explain many quantitative aspects. Among other things, it is not clear as to why the Stokes and anti-Stokes satellites have different intensities. The Raman scattering is basically a quantum effect and can be described completely by the quantum theory alone.

QUANTUM INTERPRETATION. Molecules have discrete energy levels. The Raman scattering is explained by the transitions between the vibrational levels of molecules. Suppose that a molecule absorbs an incident radiation quantum with energy $\varepsilon = \hbar\omega$. A part $\varepsilon_v = \hbar\Omega$ of the energy is absorbed by the molecule to go over to a higher vibrational level. The rest of energy is emitted as a radiation quantum having an energy $\varepsilon_1 = \varepsilon - \varepsilon_v = \hbar(\omega - \Omega)$, i.e. a frequency $\omega_1 = \omega - \Omega$. This is the mechanism behind the formation of the Stokes component. The anti-Stokes component is formed if the vibrational levels of molecules are highly excited at a certain temperature. The radiation quantum $\varepsilon = \hbar\omega$ is absorbed by a molecule on an excited vibrational level. After absorbing this quantum, the molecule returns to a lower vibrational level and the energy $\varepsilon_v = \hbar\Omega$ liberated in this case is added to the energy of the absorbed quantum. The quantum emitted in this case has an energy $\varepsilon_2 = \varepsilon + \varepsilon_v = \hbar(\omega + \Omega)$, i.e. a frequency $\omega_2 = \omega + \Omega$. This is how the anti-Stokes component is formed. It is clear that at moderate temperatures, the number of molecules participating in the emission of the Stokes components is much larger than the number of molecules capable of participating in the emission of the anti-Stokes components. This is why the intensity of the Stokes components is higher than that of the anti-Stokes components.

The difference between the intensities of the Stokes and anti-Stokes components decreases with increasing temperature because of a relative increase in the number of excited molecules capable of participating in the emission of the anti-Stokes components.

The intensity of the Raman lines is determined by the ease with which the molecules can be polarized at a certain Raman frequency, while the intensity of the absorption lines depends on the ease with which the vibrations can be excited in a molecule by the radiation of a certain frequency. These two physical factors are of different nature, and hence there is no direct relation between the intensities of the absorption lines and of the corresponding Raman scattering lines.

APPLICATIONS OF RAMAN SCATTERING. The Raman effect provides a direct method for the investigation of the molecular structure, permits measurements of their natural vibration frequencies, helps in studying the symmetry of molecules, intramolecular forces, molecular dynamics, etc. The Raman spectra characterize a molecule with such a precision that their analysis can be used for determining the composition of mixtures of molecules when the methods of chemical analysis fail to provide desired results.



Chapter **10**

Generation of Light

Basic idea:
A population inversion of energy levels makes it possible to amplify and generate light fluxes of high intensity and long coherence time.

SEC. 50. BLACKBODY RADIATION

The classical and the elementary quantum theories of the blackbody radiation are described. The properties of stimulated emission are analyzed.

RADIATION DENSITY. Light generators in everyday use are material media heated to very high temperatures. These media can be in the solid, liquid or gaseous state. We shall call all of them material bodies, or just bodies, to distinguish them from radiation which is also material object though not a body. All bodies emit, absorb and scatter electromagnetic waves, but not to the same extent. The intensity of these processes depends on the properties of the bodies, their temperature, and the frequency of electromagnetic waves. In vacuum, the electromagnetic radiation is characterized by the total density of radiation per unit volume:

$$w = \frac{1}{2}(\varepsilon_0 E^2 + B^2/\mu_0), \tag{50.1}$$

which is distributed in a certain way over the frequency spectrum. The distribution of radiation over frequencies is described by the spectral density of radiation:

$$w_\omega = \mathrm{d}w/\mathrm{d}\omega. \tag{50.2}$$

It follows from (50.2) that

$$w = \int_0^\infty w_\omega \, \mathrm{d}\omega. \tag{50.3}$$

EQUILIBRIUM DENSITY OF RADIATION. Suppose that a certain volume in space, together with the material bodies contained in it, is enclosed in an adiabatic shell. After the passage of a sufficiently long time, thermodynamic equilibrium is established between the material bodies enclosed in the adiabatic shell and the radiation contained in this volume. All the bodies will have the same temperature T, and the radiation in the cavity will have a definite spectral density which is called the equilibrium radiation density.

KIRCHHOFF'S FIRST LAW OF RADIATION. The equilibrium spectral density w_ω depends only on temperature T and is independent of the properties of bodies in the cavity, as well as of the properties of the walls of the cavity.

The validity of this statement is proved with the help of the second law of thermodynamics. Suppose that w_ω depends on the properties of the cavity and the material bodies contained in it. Let us take two different cavities at the same temperature. By hypothesis, the values of w_ω are different in these two cavities. We combine these two cavities into one. In view of different values of w_ω, an exchange of radiation energy must begin: one of the cavities must get hotter and the other colder. Thus a temperature difference will be created, which can be used for doing work. After the work has been done, the temperature difference between the cavities will be eliminated, and the system will attain a state of thermodynamic equilibrium at a temperature lower than the initial temperature since a part of the energy of the system was spent in doing work. Thus, work is done by cooling an adiabatically isolated system, which contradicts the second law of thermodynamics. Thus, the Kirchhoff first law is proved and can be written in the form

$$w_\omega = w_\omega(T). \tag{50.4}$$

ABSORPTIVE POWER AND RADIANT EMITTANCE. It is assumed that the radiation emitted by a surface obeys Lambert's law, and hence the spectral densities of radiant emittance and radiance are connected through relation (7.9a). The absorptive power A_ω is defined as the ratio of the energy absorbed by a surface element of the body per second in the frequency interval $(\omega, \omega + d\omega)$ to the entire radiant energy incident per second on this surface element in the same frequency interval. Moreover, it is assumed that the radiation falls isotropically on this surface.

KIRCHHOFF'S SECOND LAW OF RADIATION. In the state of equilibrium, the radiant energy absorbed per second by a surface element must be equal to the energy emitted during the same time by this element. This condition can be written in the form

$$\boxed{\frac{M_\omega}{A_\omega} = \frac{c}{4} w_\omega (T),} \tag{50.5}$$

in which the factor $c/4$ takes into account the connection between the energy density and the energy flux density in the case when they are distributed isotropically.

The value of this factor can be obtained from the following considerations. If the energy were transferred in one dimension and the energy flux density were distributed identically between two directions, the energy flux density in one direction would be equal to $cw_\omega/2$. However, in the case of isotropic distribution of energy flux densities in the three-dimensional case, the flux in a given direction is obtained as the sum of the projections of the flux densities in all directions onto the given direction, fluxes with positive projections only being taken into consideration. Hence the energy flux density in the direction, say, of the Z-axis, is equal to $\langle v_z \rangle w_\omega/2$, where $\langle v_z \rangle$ is the average value of the positive projection of the energy flow rate onto the Z-axis. Denoting the polar and axial angles in the spherical coordinate system by θ and φ, we obtain

$$\langle v_z \rangle = \frac{c}{2\pi} \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi/2} \cos\theta \sin\theta \, \mathrm{d}\theta = c/2,$$

where we have considered that the solid angle formed by the hemisphere is 2π. Hence we obtain the expression $cw_\omega/4$ for the energy flux density in a given direction. This expression coincides with the right-hand side of (50.5).

We direct the Z-axis in the spherical system of coordinates along the normal to the radiating surface, and denote the polar and axial angles by θ and φ. The energy flux density over the solid angle $\mathrm{d}\Omega$ is equal to $[cw_\omega/(4\pi)]\,\mathrm{d}\Omega$ in view of isotropic distribution of radiation, and the spectral density of the emissive power $\mathrm{d}P_\omega$ through a surface element $\mathrm{d}\sigma'$ perpendicular to the direction of energy flow is given by the formula $\mathrm{d}P_\omega = [cw_\omega/(4\pi)]\,\mathrm{d}\Omega\,\mathrm{d}\sigma'$. This power is emitted from an area element $\mathrm{d}\sigma$ of a luminous surface, where $\mathrm{d}\sigma' = \mathrm{d}\sigma \cdot \cos\theta$ (see Fig. 24). Hence we obtain the following expression for the spectral density of the radiant emittance:

$$M_\omega = \frac{\mathrm{d}P_\omega}{\mathrm{d}\sigma} = \int_{\Omega=2\pi} \frac{cw_\omega}{4\pi} \cos\theta \, \mathrm{d}\Omega$$

$$= \frac{cw_\omega}{4\pi} \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi/2} \cos\theta \sin\theta \, \mathrm{d}\theta = \frac{c}{4} w_\omega .$$

This is the same expression as that envisaged in (50.5), which is an expression for Kirchhoff's second law. Knowing the universal function $w_\omega(T)$, we can determine the radiant emittance from the absorbing power.

BLACKBODY. A blackbody is one which completely absorbs all the radiation incident on it, i.e. for which $A_\omega = 1$. Equation (50.5) assumes the following form for a blackbody:

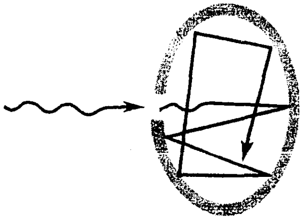


Fig. 269. A blackbody.

$$\boxed{M_\omega(T) = (c/4)\, w_\omega(T).} \qquad (50.6)$$

Thus the problem of finding $w_\omega(T)$ is reduced to the determination of the law of emission by a blackbody.

A blackbody can be obtained by making a small hole in a cavity (Fig. 269). Rays entering the cavity through the hole are almost completely absorbed due to multiple reflection and absorption at the inner walls of the cavity and hence do not leave the cavity through the hole. This situation can be practically demonstrated, say, by looking at the open windows of a house which appear dark on a bright day. The radiation coming out of the hole in the cavity can be considered as blackbody radiation to a fairly high degree of accuracy. This radiation can be used to determine M_ω and to calculate $w_\omega(T)$ with the help of formula (50.6).

Classical physics was unable to explain theoretically the shape of the function $w_\omega(T)$ measured experimentally. For sufficiently low and sufficiently high frequencies, the limiting values of $w_\omega(T)$ were explained theoretically by the Rayleigh-Jeans and the Wien laws respectively. The general relation was obtained by Planck as an interpolation relation for the limiting cases. This work laid the foundation for the development of the quantum theory.

NUMBER DENSITY OF OSCILLATION MODES. Let us consider a cavity in the form of a cube with edge L (Fig. 270). A standing wave can be formed only if a travelling wave, reflected from two opposite faces of a cube and having covered a path of length $2L$, returns to the starting point with a phase differing from that of the initial wave by $2\pi n$, where n is an integer. Without any loss of generality, it can be assumed that double reflection at the cube faces either does not change the phase at all, or changes it by 2π. Hence the condition for the formation of standing waves along each cube edge has the form

$$k \cdot 2L = 2\pi n \qquad (50.7)$$

or

$$k_x L = \pi n_x, \quad k_y L = \pi n_y, \quad k_z L = \pi n_z, \qquad (50.8)$$

where n_x, n_y, and n_z are integers.

The number $\mathrm{d}N$ of waves whose wave numbers lie between

k_x and $k_x + \mathrm{d}k_x$, k_y and $k_y + \mathrm{d}k_y$, k_z and $k_z + \mathrm{d}k_z$ is equal to the number of integers in the interval $(n_x,\ n_x + \mathrm{d}n_x)$, $(n_y,\ n_y + \mathrm{d}n_y)$ and $(n_z,\ n_z + \mathrm{d}n_z)$. Hence

$$\mathrm{d}N = \mathrm{d}n_x \mathrm{d}n_y \mathrm{d}n_z = (L/\pi)^3 \mathrm{d}k_x \mathrm{d}k_y \mathrm{d}k_z . \tag{50.9}$$

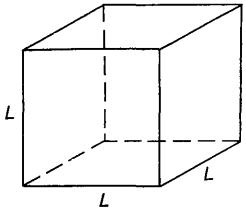


Fig. 270. To the calculation of the number of standing waves in a cubic cavity.

It is convenient to carry out calculations in the spherical system of coordinates by assuming that the segments k_x, k_y, and k_z are laid off along the axes of the Cartesian coordinate system (Fig. 271). Since k_x, k_y, and k_z are positive numbers, Eq. (50.9) assumes the following form in the spherical system of coordinates:

$$\mathrm{d}N = (L/\pi)^3 (1/8)\, 4\pi k^2 \mathrm{d}k . \tag{50.10}$$

Considering that $k = \omega/c$, we obtain the number density of standing waves:

$$\frac{\mathrm{d}N}{L^3} = \frac{1}{2}\frac{\omega^2}{\pi^2 c^3}\mathrm{d}\omega . \tag{50.11}$$

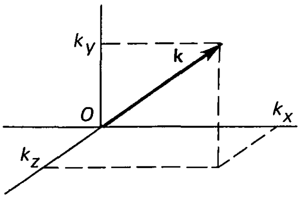


Fig. 271. Calculation of the number of standing waves in a spherical system of coordinates.

Since an electromagnetic wave can have two possible polarizations, the total number density of standing waves is twice the value given by (50.11) and is equal to

$$\frac{\mathrm{d}N_{\mathrm{tot}}}{L^3} = \frac{\omega^2}{\pi^2 c^3}\mathrm{d}\omega . \tag{50.12}$$

Each standing wave is called an oscillation mode, and the number of modes (50.12) is equal to the number of degrees of freedom of the system. If $\langle \varepsilon \rangle$ is the mean energy per degree of freedom, the energy density of standing waves is equal to

$$w_\omega(T) = \frac{\mathrm{d}N_{\mathrm{tot}}}{L^3}\langle \varepsilon \rangle = \frac{\omega^2}{\pi^2 c^3}\langle \varepsilon \rangle . \tag{50.13}$$

Thus, the determination of $w_\omega(T)$ boils down to the determination of the average energy of the oscillation modes.

RAYLEIGH-JEANS FORMULA. According to the law of equipartition of energy, each degree of freedom in a classical statistical system has an energy $kT/2$ corresponding to it. The mean kinetic energy of a harmonic oscillator is equal to its mean potential energy, and hence the average energy of a harmonic oscillator is equal to kT. This is the energy per oscillation mode. In (50.13), we put

$$\langle \varepsilon \rangle = kT . \tag{50.14}$$

This gives

$$w_\omega(T) = \omega^2 kT/(\pi^2 c^3) . \tag{50.15}$$

This relation is called the Rayleigh-Jeans formula. It was proposed in 1900 by Rayleigh (1842-1919) and substantiated in a greater detail by Jeans (1877-1946). This formula gives a very good agreement with the experiment for small values of ω. For large ω, the spectral density (50.15) considerably exceeds the observed value, and an inadmissible relation $w_\omega \to \infty$ is obtained for $\omega \to \infty$. Moreover, the total volume density of the radiation is

$$w = \int_0^\infty \omega_\omega d\omega \to \infty, \tag{50.16}$$

which is also inadmissible. The divergence of the total radiation density (50.16) is called the ultraviolet catastrophe.

WIEN'S FORMULA. It was suggested by Wien (1864-1928) that each oscillation mode is a carrier of energy $\varepsilon(\omega)$, but not all the modes of a given frequency are excited. The relative number $\Delta N/N$ of the excited modes is given by the Boltzmann distribution:

$$\Delta N/N = e^{-\varepsilon/(kT)}. \tag{50.17}$$

This leads to the following expression for the average energy per oscillation mode having a frequency ω:

$$\langle\varepsilon\rangle = \varepsilon(\omega)\Delta N/N = \varepsilon(\omega)e^{-\varepsilon/(kT)}. \tag{50.18}$$

It was concluded by Wien on the basis of general thermodynamic considerations that the energy of a mode with frequency ω is proportional to the frequency, i.e. $\varepsilon(\omega) = \hbar\omega$. The proportionality factor in this relation is given in the present-day notation as the Planck constant, which was not known at that time. Taking (50.18) into consideration, we can write (50.13) in the form

$$w_\omega(T) = \hbar\omega^3 e^{-\hbar\omega/(kT)}/(\pi^2 c^3). \tag{50.19}$$

This relation is called Wien's formula (1896) and gives a fairly good agreement with the experiment in the region of quite high frequencies. If we take the solar radiation spectrum (see Fig. 1), we can use the Rayleigh-Jeans formula for describing only those frequencies that are much lower than the one corresponding to the radiation density peak, while Wien's formula provides a correct description for frequencies that are very high and lie way beyond the peak. For a long time, the intermediate region could not be explained by any of the existing laws.

PLANCK'S FORMULA. In 1900, Planck (1858-1947) proposed an interpolation formula which is transformed into the Rayleigh-Jeans formula at low frequencies and into Wien's formula at high frequencies. This formula has the form

$$w_\omega(T) = \frac{\hbar\omega^3}{\pi^2 c^3}\frac{1}{e^{\hbar\omega/(kT)} - 1}, \tag{50.20}$$

!

The direction of propagation, polarization and phase of induced radiation coincide with the relevant characteristics of inducing radiation.
The position of maximum in the spectral density of blackbody radiation depends on the scale for which the spectral density of radiation has been determined. On the frequency scale, the maximum of the spectral density of radiation corresponds to waves longer than on the wavelength scale.

where $\hbar = 1.05 \times 10^{-34}$ J·s is called Planck's constant. For $\hbar\omega \ll kT$, this relation assumes the form (50.15), while for $\hbar\omega \gg kT$, it becomes identical with (50.19). Formula (50.20) completely describes a blackbody radiation and provides a satisfactory agreement with the experiment.

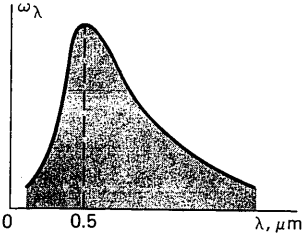


Fig. 272. Spectrum of blackbody radiation on the wavelength scale.

STEFAN-BOLTZMANN LAW. The total volume density of radiation is

$$w = \int_0^\infty w_\omega \, d\omega = \frac{k^4 T^4}{\pi^2 c^3 \hbar^3} \int_0^\infty \frac{\xi^3 d\xi}{e^\xi - 1}. \qquad (50.21)$$

Considering that

$$\int_0^\infty \frac{\xi^3 d\xi}{e^\xi - 1} = \frac{\pi^4}{15}, \qquad (50.22)$$

we can write (50.21) in the form

$$w = aT^4, \quad a = \frac{\pi^2 k^4}{15 c^3 \hbar^3} = 7.56 \times 10^{-16}\ \mathrm{J \cdot m^{-3} \cdot K^{-4}}. \qquad (50.23)$$

Usually, instead of volume density we employ the concept of radiant emittance which in accordance with (50.6) is connected with w through the relation

$$M = cw/4. \qquad (50.24)$$

Taking into account relation (50.23), we can write (50.24) in the form

$$\boxed{M = \sigma T^4,} \qquad (50.25)$$

where

$$\boxed{\sigma = ac/4 = 5.67032 \times 10^{-8}\ \mathrm{W \cdot m^{-2} \cdot K^{-4}}.} \qquad (50.26)$$

Relation (50.26) expresses the Stefan-Boltzmann law, and the constant σ is called the Stefan-Boltzmann constant.

WIEN'S DISPLACEMENT LAW. The maximum spectral density of radiation can be determined from (50.20). However, the position of the maximum depends on the scale in which the spectral density is determined. To make calculations on the wavelength scale, we must go over to wavelengths $\lambda = 2\pi c/\omega$ and to spectral density w_λ on the wavelength scale. This gives

$$w_\omega d\omega = -\, w_\lambda 2\pi c \, d\lambda/\lambda^2. \qquad (50.27)$$

Hence the spectral density distribution over wavelengths has the form (Fig. 272)

$$w_\lambda = \frac{16\pi^2 \hbar c}{\lambda^5 (\exp[2\pi c\hbar/(kT\lambda)] - 1)}. \qquad (50.28)$$

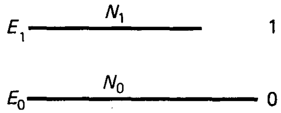


Fig. 273. Two-level system of atoms.

The maximum of the spectral density of radiation is obtained from the condition

$$\partial w_\lambda/\partial\lambda = 0, \tag{50.29}$$

which gives the following equation for defining $\lambda_{\max}$:

$$5 = xe^x/(e^x - 1), \quad x = 2\pi c\hbar/(kT\lambda_{\max}). \tag{50.30}$$

The solution of this equation is $x = 4.965$. Hence $\lambda_{\max}$ is defined by the relation

$$\boxed{\lambda_{\max}T = 2\pi c\hbar/(kx) = 0.0029 \text{ m}\cdot\text{K},} \tag{50.31}$$

called the Wien displacement law. The emission spectrum of a blackbody on the wavelength scale is shown in Fig. 272. Formula (50.31) gives a value of about 0.550 μm for $\lambda_{\max}$ in the emission spectrum of the Sun ($T_{\text{sur}} \simeq 6000$ K). To calculate the maximum from the frequency spectrum, we must solve the equation $\partial w_\omega/\partial\omega = 0$ instead of (50.29). As a result, we obtain a value of about 2.1×10^{15} s^{-1} for $\omega_{\max}$ in the emission spectrum of the Sun. This corresponds to a wavelength of about 0.88 μm.

ELEMENTARY QUANTUM THEORY. It is not possible to derive Planck's formula from classical considerations. However, it can be substantiated with the help of elementary arguments based on the quantum concepts of emission and absorption of light. Although these arguments do not lead to a complete quantitative solution of the problem on the blackbody radiation, they do provide a clear idea about the mechanism of dynamic equilibrium between radiation and material bodies.

A hollow cavity and radiation in it are in thermodynamic equilibrium. The radiation is represented as an aggregate of quanta with the energy $\varepsilon = \hbar\omega$. The quanta may be absorbed by atoms which go over to a higher energy level with the energy $E_1 = E_0 + \hbar\omega$, where E_0 is the initial energy level of the atom. A transition of the atom from level E_1 to level E_0 is accompanied by the emission of a quantum with the energy $E_1 - E_0 = \hbar\omega$. Let us denote these levels by 0 and 1 (Fig. 273) and call them lower and upper levels respectively.

Dynamic equilibrium is attained with the help of a constant exchange of quanta between the radiation field and material bodies, resulting in transitions of atoms between energy levels and a change in the number of quanta in the radiation field. In accordance with the detailed balancing principle, the exchange of quanta must be balanced separately for each frequency (to be more precise, for each radiation mode). Hence we shall consider just one frequency. For other frequencies, the same

?

Is the calculation of the number density of oscillation modes connected with the concretization of boundary conditions? Formulate the conditions in which the Rayleigh-Jeans and the Wien formulas are obtained. For which scale of spectral energy density is the Wien displacement law formulated?

arguments are applicable, and only the levels between which dynamic equilibrium is established are different.

SPONTANEOUS AND INDUCED TRANSITIONS. In accordance with the law of conservation of energy, transitions from the lower to the upper level are possible only through the absorption of an energy quantum, i. e. through the radiation incident on the atom. Such transitions are called induced transitions. Transitions from the upper to the lower level can be either induced by the radiation incident on the atom, or occur spontaneously, irrespective of the radiation incident on the atom.

Let us denote by $\nu_{10}^{(s)}$ the frequency of spontaneous atomic transitions from the upper to the lower level, by $\nu_{10}^{(i)}$ the frequency of induced transition from the upper to the lower level, and by $\nu_{01}^{(i)}$ the frequency of induced transition from the lower to the upper level. The condition for dynamic equilibrium can be written in the form

$$\nu_{10}^{(s)} + \nu_{10}^{(i)} = \nu_{01}^{(i)}. \tag{50.32}$$

EINSTEIN'S COEFFICIENTS. In 1917, Einstein (1879-1955) proposed an elementary quantum concept of dynamic equilibrium between radiation and material medium. This concept provided a correct formula for the blackbody radiation.

Let us denote by A_{10} the probability of spontaneous transition $1 \to 0$ per second, and by N_1 the number density of atoms on the upper level. By the definition of these quantities, we can write

$$\nu_{10}^{(s)} = N_1 A_{10}. \tag{50.33}$$

Obviously, the frequency of induced transitions is proportional to the radiation density. We denote by B_{10} and B_{01} the probabilities of induced transitions $1 \to 0$ and $0 \to 1$ per second, referred to the spectral radiation density w_ω, and by N_0 the number density of atoms on the lower level. By definition of these quantities, we can write

$$\nu_{10}^{(i)} = N_1 w_\omega B_{10}, \quad \nu_{01}^{(i)} = N_0 w_\omega B_{01}. \tag{50.34}$$

Taking (50.33) and (50.34) into account, we can write (50.32) in the form

$$N_1 A_{10} + N_1 w_\omega B_{10} = N_0 w_\omega B_{01}. \tag{50.35}$$

In the equilibrium state, Boltzmann's distribution is applicable. This distribution has the following form for the number density of atoms:

$$N_1 = A\mathrm{e}^{-E_1/(kT)}, \quad N_0 = A\mathrm{e}^{-E_0/(kT)}, \tag{50.36}$$

where A is the normalization constant. Substituting (50.36) into (50.35) and cancelling out the common factor A in the obtained expression, we obtain

$$A_{10}e^{-E_1/(kT)} + w_\omega B_{10}e^{-E_1/(kT)} = w_\omega B_{01}e^{-E_0/(kT)}. \qquad (50.37)$$

The quantities A_{10}, B_{10} and B_{01} are called the Einstein coefficients.

It follows from physical considerations that as $T \to \infty$, w_ω must also tend to infinity. Dividing both sides of (50.37) by w_ω, we get

$$B_{10} = B_{01}. \qquad (50.38)$$

Hence relation (50.37) can be written in the form

$$w_\omega = \frac{A_{10}}{B_{10}} \frac{1}{e^{\hbar\omega/(kT)} - 1}, \qquad (50.39)$$

where $\hbar\omega = E_1 - E_0$. Within the framework of the elementary quantum theory, it is impossible to determine the ratio A_{10}/B_{10} theoretically. This quantity is determined with the help of a rigorous quantum theory of radiation. However, this ratio can be determined if we consider that at low frequencies formula (50.39) must coincide with Rayleigh-Jeans formula (50.15), while at high frequencies it should be transformed into Wien's formula (50.19).

For $\hbar\omega \ll kT$, we can assume that $\exp[\hbar\omega/(kT)] \approx 1 + \hbar\omega/(kT)$. This allows us to write (50.39) in the form

$$w_\omega = \frac{A_{10}}{B_{10}} \frac{kT}{\hbar\omega}. \qquad (50.40)$$

A comparison of (50.40) and (50.15) gives

$$A_{10}/B_{10} = \hbar\omega^3/(\pi^2 c^3). \qquad (50.41)$$

Formula (50.39) in this case assumes the form

$$w_\omega = \frac{\hbar\omega^3}{\pi^2 c^3} \frac{1}{e^{\hbar\omega/(kT)} - 1}, \qquad (50.42)$$

which is identical with Planck's formula (50.20).

Although the elementary quantum theory of the blackbody radiation does not allow a theoretical computation of Einstein's coefficients, it does demonstrate the necessity for the existence of spontaneous and induced transitions. The important relation (50.38) is satisfied for the probabilities of induced transitions.

The quanta emitted as a result of spontaneous transitions propagate in random directions, have random polarizations and random phases. Quanta emitted as a result of induced transitions correlate in their properties with the radiation responsible for the transition. The induced radiation has the same polarization, direction of propagation and the phase as the radiation inducing the transition. This property of induced radiation is extremely important for its applications.

SEC. 51. OPTICAL AMPLIFIERS

Conditions for the amplification of a light flux passing through a medium are discussed.

PASSAGE OF LIGHT THROUGH A MEDIUM. When light passes through a medium, quanta are exchanged between the light beam and the atoms of the medium through induced transitions and spontaneous emission of quanta takes place. Let us denote the emission frequency and the number density of atoms in the upper and lower levels by ω, N_1, and N_0 respectively (see Fig. 273). The spectral density of radiation of frequency ω is denoted by w_ω. This quantity changes due to the induced absorption of quanta by the atoms of the medium, leading to a decrease in the flux density, and due to the induced emission by the atoms, which causes an increase in the density w_ω. On the basis of (50.34), the law of energy conservation in induced transitions can be written in the form

$$\mathrm{d}w_\omega/\mathrm{d}t = \hbar\omega w_\omega B_{10}N_1 - \hbar\omega w_\omega B_{01}N_0 = \hbar\omega w_\omega B(N_1 - N_0), \tag{51.1}$$

where $B = B_{10} = B_{01}$. Using the notation

$$\alpha = \hbar\omega B(N_1 - N_0)/v \tag{51.2}$$

we can write (51.1) in the form

$$\mathrm{d}w_\omega/\mathrm{d}t = \alpha v w_\omega, \tag{51.3}$$

where v is the velocity of light of frequency ω in the medium. This equation can be written in the form

$$\mathrm{d}S/\mathrm{d}t = \alpha v S, \tag{51.4}$$

where

$$S = v w_\omega \tag{51.5}$$

is the energy flux density. Assuming that light propagates in the direction of the Z-axis, we can write

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d}S}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}S}{\mathrm{d}z}v, \tag{51.6}$$

where $\mathrm{d}z/\mathrm{d}t = v$ is the velocity of light propagation in the medium. Obviously, we are speaking in this case of the group velocity with which the energy propagates. Taking (51.6) into consideration, we can write Eq. (51.4) in the form

$$\mathrm{d}S/\mathrm{d}z = \alpha S. \tag{51.7}$$

This equation has the solution

$$S(z) = S_0 e^{\alpha z}, \tag{51.8}$$

where $S_0 = S(0)$.

BURGER'S LAW. The exponential dependence (51.8) of the flux density on distance is called Burger's law. In the state of thermodynamic equilibrium, the number densities of atoms are described by the Boltzmann distribution. It follows from the inequality $E_1 > E_0$ that $N_1 < N_0$ and hence $\alpha < 0$. This means that as light passes through a medium, its flux density decreases. The decrease in the flux density takes place according to the following mechanism. On account of the induced transitions of atoms from the lower energy level to the upper one, the energy flux density decreases. During transitions of the atoms from the upper energy level to the lower one, only a part of the quanta rejoin the flux, viz. the quanta emitted as a result of induced transitions. Spontaneously emitted quanta do not return to the flux, and hence the energy flux density decreases as light passes through the medium.

AMPLIFICATION CONDITIONS. If a system of atoms is brought into a nonequilibrium state and the Boltzmann distribution is quite strongly violated as a result, the number density of atoms on different levels can be changed in such a way that $N_1 \approx N_0$ or even $N_1 > N_0$. In the former case, $\alpha \simeq 0$ and the beam would propagate through the medium without being absorbed. In the latter case, $\alpha > 0$ and the beam passing through the medium would be amplified, i. e. the medium would act as an amplifier for the light flux.

This makes it possible to create generators and amplifiers of waves on the basis of induced emission. These amplifiers are called lasers for the radiation in the optical range, while for microwave radiation they are termed masers. For fundamental research in the field of quantum electronics leading to the creation of new types of generators and amplifiers, viz. masers and lasers, the Soviet physicists Prokhorov (b. 1916) and Basov (b. 1922), as well as the American scientist Townes (b. 1915), were awarded the Nobel Prize in physics in 1964. The idea of possible application of induced emission for amplification of light flux was put forth in 1940 by the Soviet physicist V. A. Fabrikant (b. 1907).

EFFECT OF LIGHT FLUX ON THE POPULATION DENSITY OF LEVELS. By inducing transitions between atomic energy levels, a light beam causes a change in their population densities. Let us denote by N the total number density of the atoms. By definition,

$$N = N_1 + N_0. \tag{51.9}$$

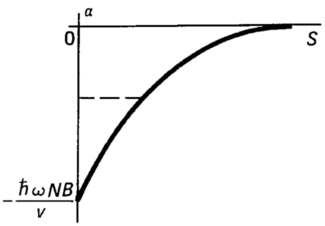


Fig. 274. Dependence of the attenuation factor on the energy flux in a light beam.

The population density of the upper level changes as a result of induced and spontaneous transitions and can be described by the equation

$$\mathrm{d}N_1/\mathrm{d}t = -w_\omega B(N_1 - N_0) - N_1/\tau, \tag{51.10}$$

where N_1/τ takes into account the frequency of spontaneous transitions. Substituting the expression for N_0 from (51.9) into (51.10), we obtain

$$\mathrm{d}N_1/\mathrm{d}t = -SB(2N_1 - N)/v - N_1/\tau, \tag{51.11}$$

where $w_\omega = S/v$. When stationary state is attained, the following condition must be satisfied:

$$\mathrm{d}N_1/\mathrm{d}t = 0. \tag{51.12}$$

Taking into account Eq. (51.11), we can write this condition in the form

$$-SB(2N_1 - N)/v - N_1/\tau = 0. \tag{51.13}$$

This gives

$$N_1 = \frac{N}{2}\,\frac{1}{1 + v/(2\tau BS)}. \tag{51.14}$$

It can be seen from this equation that the population density of the upper level increases with increasing energy flux density S. This leads to a corresponding change in the value of the coefficient

$$\begin{aligned}\alpha &= \hbar\omega B(N_1 - N_0)/v = \hbar\omega B(2N_1 - N)/v \\ &= -\frac{\hbar\omega NB}{v}\left[\frac{1}{1 + 2\tau BS/v}\right].\end{aligned} \tag{51.15}$$

SATURATION CONDITIONS. The dependence of α on S is shown in Fig. 274. It can be seen that absorption takes place for all values of S. As $S \to \infty$, the absorption completely stops and saturation sets in. Since α appears in (51.8) in the exponential term, it is assumed that saturation practically sets in when α becomes equal to twice its minimum value, i. e. when the denominator in (51.15) is doubled. Hence the saturation condition can be written in the form

$$v/(2\tau BS) = 1. \tag{51.16}$$

Consequently, saturation takes place at the energy flux density

$$\boxed{S = v/(2\tau B).} \tag{51.17}$$

Thus, the light flux levels out the population density of two levels between which it causes induced transitions, and may even equalize them approximately for quite large flux densities.

?

Under what conditions does the light passing through a medium (a) gets attenuated? (b) passes without an attenuation? (c) gets amplified? How are these conditions realized? Describe the three-level mechanism of creating population inversion of levels. What requirements are imposed on energy levels?

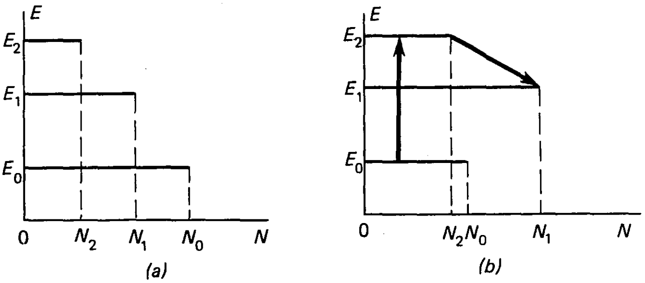


Fig. 275. Creation of population inversion in a three-level system: (*a*) normal population, (*b*) inverse population.

However, it cannot cause a population inversion between these levels.

CREATION OF POPULATION INVERSION. Population inversion of energy levels can be caused by a certain action on the atoms, irrespective of the light being amplified.

The simplest method of creating population inversion is realized in three-level systems (Fig. 275). The distribution of population densities in the equilibrium state of a system is shown in Fig. 275*a*. When high-power radiation of frequency $\omega_p = (E_2 - E_0)/\hbar$ acts on the system, the population densities of the levels E_0 and E_2 practically equalize if condition (51.17) is satisfied. We assume that the lifetime of the atoms on the level E_2 is very small and they spontaneously go over to the level E_1 on which they have a sufficiently long lifetime. Obviously, atoms will get accumulated on the level E_1, and hence a population inversion between the levels E_1 and E_0 will take place (Fig. 275*b*). This transition can be used for amplifying light with the frequency $\omega = (E_1 - E_0)/\hbar$.

SEC. 52. LASERS

The principle of operation and properties of lasers are described.

SCHEMATIC DIAGRAM OF A LASER. A medium with population inversion density, capable of amplifying the light flux passing through it, is called an active medium. Let us fill the space between the plates of a Fabry-Perot interferometer (Fig. 276) with an active medium. A light flux passing through the active medium will be amplified between successive reflections from the mirrors. This system forms an active optical resonator. The flux amplification during a passage through the active medium takes place in accordance with formula (51.8). The radiation is partially attenuated as a result of reflection at the mirrors. One of the mirrors has the highest possible reflectivity, while the other mirror transmits a certain

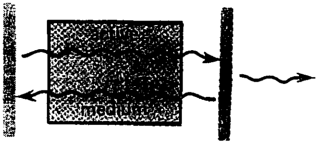


Fig. 276. Schematic diagram of a laser.

portion of light forming the radiation called laser radiation. Besides reflection at the mirrors, other factors like scattering in the medium and diffraction effects are also responsible for the losses. For the system to work as a generator of light, a certain balance must be ensured between the amplification of the light flux passing through the active medium and the attenuation of radiation due to all factors, including the laser radiation itself.

LASING THRESHOLD. The basic operating cycle for a laser involves two successive passages of light through the active medium and the corresponding reflections at the mirrors. The energy losses can be taken into account by using the effective reflectivities ρ_1 and ρ_2 of the mirrors. These coefficients take into account not only the losses due to reflection at the mirrors (these reflections are generally different since the laser radiation emerges from one of these mirrors), but also other types of losses mentioned above. Hence ρ_1 and ρ_2 are smaller than the values of reflectivities of the resonator mirrors alone. Light is reflected twice during one cycle of operation, and hence the flux attenuation is proportional to $\rho_1\rho_2$. In the active medium, light covers a distance of $2L$ during one cycle. Hence, in accordance with (51.8), the flux amplification per cycle is proportional to $\exp(\alpha 2L)$, where α is the gain factor (51.2) per cycle. The total amplification of the energy flux density during one cycle is given by the formula

$$S = S_0\rho_1\rho_2 e^{2\alpha L}, \tag{52.1}$$

where S_0 is the energy flux density at the beginning of the cycle. (Any instant of time can be chosen as the beginning of the cycle.) Equation (52.1) can be written in the form

$$S = S_0 e^{2\alpha L - 2f}, \tag{52.2}$$

where

$$2f = -\ln(\rho_1\rho_2). \tag{52.3}$$

It can be seen from (52.2) that lasing begins when the energy acquired by the light flux in the active medium per cycle exceeds the energy losses, including the energy of the laser radiation leaving the system. At the lasing threshold, the energy flux density in the system is not so high as to change population density of the levels and hence we can assume that $\alpha = \alpha_0$ for the lasing threshold, where α is the gain factor in the absence of a light flux. Consequently, the lasing threshold condition can be written in the following form in accordance with (52.2):

$$\alpha_0 L = f. \tag{52.4}$$

STEADY-STATE LASING CONDITIONS. In the steady-

!

Lasing begins only when the energy transferred per cycle from the active medium in the resonator to the light flux becomes higher than the total losses of the light flux in the resonator, including the energy carried away by the laser radiation. The beginning of lasing is characterized quantitatively by the lasing threshold condition.

?

What is the physical meaning of the lasing threshold and of the steady-state generation condition?

Describe the Q-switching method for increasing the output power.

state generation, the energy losses are compensated for by the energy obtained by the light beam from the active medium. Steady-state generation can be achieved for quite substantial energy flux densities, hence we should choose for α its value in the presence of flux. The steady-state lasing condition has the form

$$\alpha L = f. \tag{52.5}$$

Q-FACTOR. Since a laser is an optical resonator, the terminology of the oscillation theory is widely used in the theory of lasers. The loss of energy per oscillation period is usually characterized by the Q-factor, viz, the ratio of the energy W stored in the system to the energy losses ΔW per cycle:

$$Q = W/\Delta W. \tag{52.6}$$

The energy stored in the system is given by

$$W = w\sigma L, \tag{52.7}$$

where σ is the cross-sectional area of the laser beam. The energy lost during one cycle is given by

$$(w/2)\,\sigma L(1 - e^{-2f}), \tag{52.8}$$

where we have taken into account the fact that the volume density w of energy is the sum of two equal densities of energy fluxes moving in opposite directions. Considering that the duration of one cycle is equal to $2L/v$ and denoting the period of laser radiation by $T = 2\pi/\omega$, we find that the energy lost during one cycle is given by the formula

$$\Delta W = \frac{(w/2)\,\sigma L(1 - e^{-2f})}{2L/v}\,T = (1/2)\,wv\sigma fT, \tag{52.9}$$

where $\exp(-2f) \simeq 1 - 2f$ since $f \ll 1$. Hence, we obtain from formula (52.6)

$$Q = \frac{\sigma wL}{wv\sigma fT/2} = \frac{2L}{f\lambda} = \frac{m}{f}, \tag{52.10}$$

where $m = 2L/\lambda = L/(\lambda/2)$ is the number of standing half-waves in the resonator, and λ is the radiation wavelength. With the help of (52.10), condition (52.4) can be expressed in a very clear form:

$$\alpha_0\lambda/2 = 1/Q, \tag{52.11}$$

i.e. the necessary condition for lasing is that the amplification over a half wavelength be equal to (or larger than) the reciprocal of the Q-factor.

The lower the Q-factor, the higher the lasing threshold. Hence the lasing threshold is attained earlier for the axial rays, and the energy of the laser radiation is predominantly carried by these rays. This means that the laser radiation is concentrated over a narrow parallel beam of rays, the angle of divergence of this beam being determined mainly by diffraction effects.

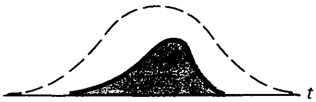


Fig. 277. Pulsed pumping of a laser (the dashed curve shows pumping pulse while the solid curve corresponds to laser pulse).

CONTINUOUS-WAVE (CW) AND PULSED LASERS. The creation of population inversion of levels is called pumping. Pumping of lasers can be carried out by quite diverse means and will be considered in the context of particular types of lasers. According to the nature of its time dependence, pumping may be continuous or pulsed. If pumping is carried out by pulses, the laser radiation also will be pulsed. After the onset of a pumping pulse, the population density of levels begins to change. When the lasing threshold conditions (52.4) are satisfied, laser radiation begins to be emitted (Fig. 277).

Under continuous pumping, which ensures that condition (52.5) is satisfied at all instants, the laser radiation is continuous. However, it must be noted that pulsed emission is also possible under continuous pumping, as well as the emission of several radiation pulses for a single pumping pulse.

ENHANCEMENT OF THE EMISSION POWER. To increase the power of laser radiation, it is necessary to increase the number of atoms participating in the amplification of the light flux in the laser resonator through induced emission, and to decrease the pulse duration.

Q-SWITCHING METHOD. To increase the number of atoms participating simultaneously in the amplification of light flux, we must delay the onset of lasing to enable the accumulation of the largest possible number of excited atoms causing the population inversion. For this purpose, it is necessary to raise the lasing threshold and to reduce the Q-factor. This can be done by increasing the losses in the light flux. For example, if the parallelism of the mirrors is violated, the Q-factor of the system sharply decreases. If pumping is started under such conditions, lasing does not begin even for a significant population inversion of levels, since the lasing threshold is quite high. The turning of a mirror to make it parallel to the other one increases the Q-factor of the system and thus lowers the lasing threshold. When the Q-factor of the system ensures the onset of lasing, the population inversion of the levels will be quite significant. Hence the laser output power will be considerably increased. This method of controlling the laser output power is called the Q-switching method.

The duration of a laser pulse depends on the time during which the population inversion changes due to the emission of a pulse to such an extent that the system no longer satisfies the lasing condition. Pulse duration depends on many factors but is usually of the order of 10^{-7}-10^{-8} s.

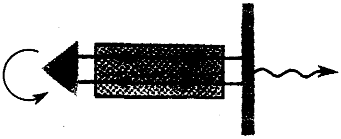


Fig. 278. Q-switching by a rotating prism.

The switching of Q-factor with the help of a rotating prism is used quite widely (Fig. 278). At a certain position, the prism ensures a total backward reflection of a ray incident along the resonator axis. The prism rotates at a frequency of several tens or hundreds of hertz. The laser radiation pulses have the same frequency, although their duration is 10^{-6}-10^{-8} s.

A higher pulse repetition frequency can be attained by switching the Q-factor with the help of a Kerr cell (see Sec. 45). The Kerr cell and a polarizer are placed in a resonator. The polarizer ensures the lasing of only a certain polarization, while the Kerr cell is oriented in such a way that when a voltage is applied to it, light with this polarization does not pass through it. During laser pumping, the voltage applied to the Kerr cell is removed at an instant when the lasing started in the resonator is the strongest.

There are other methods also in which losses can be introduced in order to attain Q-switching.

SEC. 53. LASER RADIATION

Properties of laser radiation are described.

RADIATION MODES. By the definition of steady state, the steady-state laser radiation must form standing waves in optical resonator. In its most general form, a standing wave can be represented as a superposition of elementary standing waves called the oscillation modes. Laser radiation corresponding to these oscillation modes are called the radiation modes of a laser.

An oscillation mode depends on the geometrical properties of the resonator, the refractive index of the active medium, and generally on the boundary conditions of the resonator. For the sake of definiteness, let us consider a resonator with plane rectangular mirrors, and a cylindrical resonator with spherical mirrors.

RESONATOR WITH PLANE RECTANGULAR MIRRORS (Fig. 279). The partial solution of wave equation (2.9) has the form

$$E = E_0 \begin{bmatrix} \sin \omega t \\ \cos \omega t \end{bmatrix} \begin{bmatrix} \sin k'_x x \\ \cos k'_x x \end{bmatrix} \begin{bmatrix} \sin k'_y y \\ \cos k'_y y \end{bmatrix} \begin{bmatrix} \sin k'_z z \\ \cos k'_z z \end{bmatrix}. \tag{53.1}$$

An arbitrary combination of factors on the right-hand side can be chosen (one from each bracket). The choice of the combi-

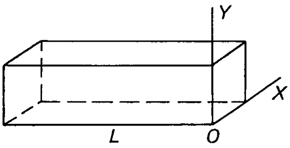


Fig. 279. A rectangular optical resonator.

nation depends on the boundary conditions that have to be satisfied. In this equation, E_0 is the constant amplitude defining the polarization of the oscillations. Substitution of the combination of factors chosen in (53.1) into Eq. (2.9) leads to the equation

$$k_x'^2 + k_y'^2 + k_z'^2 = k'^2 = \mu\varepsilon\omega^2, \tag{53.2}$$

where the wave number k' corresponds to the medium, i.e. $k' = 2\pi/\lambda_m$, λ_m being the wavelength in the medium. In Eq. (2.9), we have made the substitution $\varepsilon_0 \to \varepsilon$ and $\mu_0 \to \mu$ to write this equation for the medium. Considering that $\mu_\varepsilon = 1/v^2$, we obtain from (53.2)

$$k' = \omega/v = \omega n/c = kn, \tag{53.3}$$

where $n = c/v$ is the refractive index of the medium, and $k = \omega/c$ is the wave number in vacuum.

Standing waves emerge in each dimension of the rectangular resonator if the cyclicity condition is met:

$$m\lambda_m = 2l, \tag{53.4}$$

where λ_m is the wavelength in the medium and l is the size of the resonator in the dimension under consideration. A standing wave can be presented as the sum of two travelling waves propagating in opposite directions. The cyclicity condition means that after two successive reflections at the mirrors and the passage of two lengths of the resonator, all the characteristics of a wave must resume their initial values. In particular, this means that these waves must have a definite polarization. These waves cannot be nonpolarized. The nature of their polarization depends on the resonator. Strictly speaking, we must introduce into the right-hand side of (53.4) the variation of the phase of the waves as a result of reflection at the mirrors. However, this circumstance is not of any significance for the analysis of the process, and simply makes the formulas more cumbersome. Therefore, we shall not take it into consideration.

Assuming that the Z-axis is perpendicular to the mirror surfaces, we can write condition (53.4) in the form

$$\begin{aligned} k_x'a &= m_x\pi, \\ k_y'b &= m_y\pi, \\ k_z'L &= m_z\pi, \end{aligned} \tag{53.5}$$

where a and b represent the length of the resonator along X and Y-axes respectively.

It should be noted that conditions (53.5) have been obtained not in the form of the boundary conditions that must be always satisfied for obtaining a solution of the problem formulated as a set of differential equations. These conditions have been obtained as a corollary of a more general physical requirement rather than the mathematical requirement (53.4) of the cyclicity of the wave field. This is quite important, since this requirement explicitly contains the assumption that a standing wave is formed as a result of the superposition of two travelling waves.

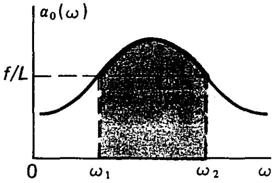


Fig. 280. Method of determining the frequency of generated modes.

AXIAL (LONGITUDINAL) MODES. Modes are characterized by a set of numbers (m_x, m_y, m_z). The mode $(0, 0, m_z)$ is called the principal mode. It has no nodal points in the plane perpendicular to the Z-axis and describes a standing wave as a superposition of two travelling waves moving towards each other in a direction parallel to the Z-axis. Outside the resonator, this standing wave has a corresponding wave travelling parallel to the laser axis. In the theory of waveguides, this mode is called the transverse electromagnetic mode and is denoted by TEM_{00m_z}. For the radiation frequencies corresponding to this mode, we obtain the following expression by combining (53.3) with (53.5):

$$\omega_{m_z} = k'_z c/n = c\pi m_z/(Ln). \tag{53.6}$$

The possible cyclic frequencies of the axial modes are separated by the interval

$$\Delta\omega = c\pi \cdot \Delta m_z/(Ln) = c\pi/(Ln), \tag{53.7}$$

since the number m_z for neighbouring frequencies differs by unity ($\Delta m_z = 1$). The frequency separation is $\Delta\nu = \Delta\omega/(2\pi) = c/(2Ln)$. Considering that $\nu = c/\lambda$, we can write this relation in the form $\Delta\nu/\nu = \lambda/(2Ln)$.

A laser generates only those frequencies that satisfy the lasing threshold conditions. The dependence of α_0 on ω is shown in Fig. 280. Lasing is possible only in the frequency interval between ω_1 and ω_2. The frequencies generated by the laser are separated by equal distances (53.7) from one another. The number of frequencies falling in the interval between ω_1 and ω_2 depends on the specific conditions. For the helium-neon laser ($\lambda = 632.8$ nm), the number of such frequencies varies from 5 to 10, for a ruby laser, there are a few hundred of them, while a dye laser has several hundreds of such frequencies. However, it is possible to establish a unimode or two-mode operation as well.

WIDTH OF EMISSION LINES. Each of the emitted lines in the amplification band is not strictly monochromatic and has a finite width. The width $\Delta\nu$ of the line can be determined from the general condition

$$\Delta\nu\tau \simeq 1, \tag{53.8}$$

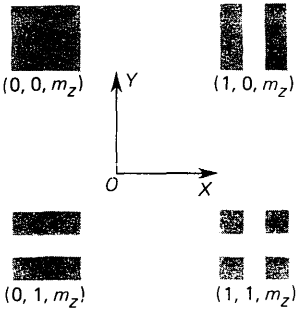


Fig. 281. Laser radiation modes for a rectangular resonator.

where τ is the emission time. For Q-switched lasers, $\tau \simeq 10^{-8}$ s, and hence $\Delta\nu = 10^8$ Hz. In the CW-regime, any value can be chosen for the emission time τ, and hence the linewidth can be decreased theoretically to any extent. In practice, however, uncontrollable fluctuations of the refractive index, variations in the value of L due to vibrations, thermal expansion of the frame due to temperature variations, etc. lead to a broadening of the lines. For example, it follows from (53.6) that $\Delta\nu/\nu = -\Delta L/L$. Hence a slight variation $\Delta L = \lambda/50$ for $L = 1$ m and $\lambda = 0.5$ μm gives $\Delta\nu/\nu = 10^{-8}$. However, all these factors do not play a significant role and their effect can be minimized by improving the technical aspects. The factors that are of principle importance are the Brownian motion in the mirrors and the spontaneous emission by the atoms of the medium. The broadening due to these factors is very small, though, and has an order of magnitude $\Delta\nu \simeq 10^2$-10^{-1} Hz. For $\nu \simeq 10^{15}$ Hz, this gives $\Delta\nu/\nu = 10^{-13}$-10^{-16}. At the present time, this limit has actually been achieved.

SIDE MODES. The values of m_x and m_y characterize the number of nodal points for a standing wave in the XY plane. For $m_x = 0$ and $m_y = 0$, there are no nodal points and hence the axial mode (0, 0, m_z) emerging from the laser gives an intensity distribution without any nodes (Fig. 281). Modes with $m_x \neq 0$ and $m_y \neq 0$ are called side modes. The intensity distribution of the radiation corresponding to these modes is characterized by the zero-intensity lines corresponding to the nodal points of the standing waves for $m_x \neq 0$ and $m_y \neq 0$. The intensity distribution of the radiation emerging from the laser for the first few side modes is shown in Fig. 281. Taking (53.5) into account, we obtain from (53.2)

$$\omega_{m_x, m_y, m_z} = \frac{\pi c}{n}\sqrt{\left(\frac{m_x}{a}\right)^2 + \left(\frac{m_y}{b}\right)^2 + \left(\frac{m_z}{L}\right)^2}. \tag{53.9}$$

We denote by $\Delta\omega_x$ and $\Delta\omega_z$ the variation of frequency upon a transition from one mode to another for $\Delta m_x = 1$ and $\Delta m_z = 1$. From (53.9), we have

$$\Delta\omega_x = \frac{\pi c}{n}\left[\left(\frac{m_x}{a}\right)^2 + \left(\frac{m_y}{b}\right)^2 + \left(\frac{m_z}{L}\right)^2\right]^{-1/2}\frac{m_x}{a^2}, \tag{53.10}$$

$$\Delta\omega_z = \frac{\pi c}{n}\left[\left(\frac{m_x}{a}\right)^2 + \left(\frac{m_y}{b}\right)^2 + \left(\frac{m_z}{L}\right)^2\right]^{-1/2}\frac{m_z}{L^2}. \tag{53.11}$$

For $m_x = m_y = 0$, (53.11) yields formula (53.7). Dividing the left- and right-hand sides of (53.10) by (53.11) termwise, we

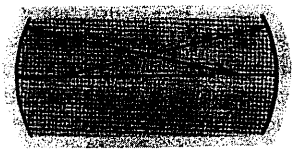


Fig. 282. A cylindrical resonator with spherical mirrors.

obtain

$$\frac{\Delta\omega_x}{\Delta\omega_z} = \frac{m_x}{m_z}\left(\frac{L}{a}\right)^2. \tag{53.12}$$

Considering that $m_z \propto L/\lambda$, and $m_x \simeq 1$, we get

$$\frac{\Delta\omega_x}{\Delta\omega_z} \propto \frac{\lambda}{a}\frac{L}{a} \leqslant 1. \tag{53.13}$$

This means that the separation between the side mode frequencies is much smaller than between the axial mode frequencies.

CYLINDRICAL RESONATOR WITH SPHERICAL MIRRORS. For standing waves in this resonator, the surfaces of the mirrors are constant-phase surfaces. In other words, the wave front varies along the Z-axis and when it reaches the mirrors, it coincides with the mirror surface (Fig. 282). If the mirrors have the same radius of curvature, the wave front at the centre of the resonator will be plane. As usual, the standing wave can be assumed as a superposition of two waves propagating in opposite directions. In one cycle during which the wave is reflected twice at the mirrors and passes twice through the resonator, all the characteristics of the reflected and transmitted waves must assume their initial values. Calculations show that the cyclicity condition for the reflected waves has the form

$$[m_z + (1/\pi)(2p + q + 1) \arccos (1 - L/p)] \lambda_{\mathrm{m}} = 2L, \tag{53.14}$$

where m_z, p and q are integers characterizing a mode. For an electric field, the solution can be written in cylindrical coordinates with the axis coinciding with the Z-axis in the following form:

$$E(r, \varphi, z) = E_0 \left[\frac{r}{\sigma}\sqrt{2}\right]^q L_q^p\left(2\frac{r^2}{\sigma^2}\right) \exp \times \left[-\frac{r^2 \cos q\varphi}{\sigma^2}\right] \sin kz. \tag{53.15}$$

Here, L_q^p are Laguerre polynomials and σ is defined by the relation

$$\sigma^4 = \left(\frac{\lambda}{\pi}\right)^2 \frac{R^2 L}{(2R - L)}, \tag{53.16}$$

R being the radius of curvature of the mirrors.

Transverse electromagnetic waves denoted by TEM_{pqm_z} are the most significant ones. For $p = 0$ and $q = 0$, the mode is axial. Condition (53.14) for the axial mode coincides with

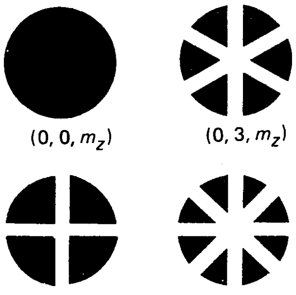


Fig. 283. Laser radiation modes in a cylindrical resonator with spherical mirrors.

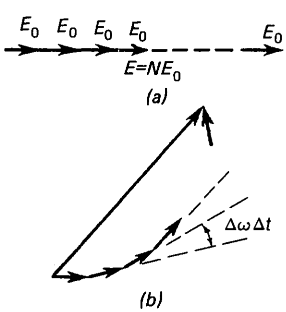


Fig. 284. Addition of amplitudes during mode synchronization.

!

The concentration of laser radiation in one direction in a resonator with plane mirrors is due to the fact that the Q-factor is maximum in this direction, while the lasing threshold is minimum. The divergence of a laser beam is mainly caused by diffraction phenomena.

condition (53.4) for the axial mode in a rectangular resonator. The modes with $p \neq 0$ and $q \neq 0$ are side modes. The value of p determines the number of nodal points along the radius, while the value of q is equal to half the number of axial nodal points. Figure 283 shows the intensity distribution of the radiation of various modes without any nodal points along the radius ($p = 0$) at the output of a laser. For $p = 1$ and $q = 0$, a node appears at the point $r = 0$. In the mode (0, 0, m_z) in Fig. 283, this node can be shown by a dark spot. For other values of p, concentric circles of zero intensity are observed.

MODE SYNCHRONIZATION. The pulse duration of laser radiation obtained by Q-modulation usually varies between 10^{-7} and 10^{-8} s. With the help of Kerr cells, pulses of duration $\tau = 10^{-9}$ s can be obtained. Still shorter pulses ($\sim 10^{-12}$ s) can be obtained as a result of mode synchronization. The essence of mode synchronization lies in the following.

In the multimode operation, a laser emits a spectrum of frequencies which are separated for axial modes by $\Delta\omega$ defined by formula (53.7). If all the modes are added in the same phase, the amplitude of the resultant wave will be the sum of the mode amplitudes (Fig. 284*a*). After an interval of time Δt, the phases of the waves change by $\Delta\omega\Delta t$, $2\Delta\omega\Delta t, \ldots, (N-1)\Delta\omega\Delta t$, where N is the number of modes participating in the formation of the resultant amplitude. The pattern of summation of complex amplitudes in an interval Δt is shown in Fig. 284*b*. Upon a further increase in Δt, the amplitude decreases and then begins to increase. The initial situation in which the amplitudes are added in the same phase is repeated after a time T satisfying the obvious condition

$$\Delta\omega T = 2\pi. \tag{53.17}$$

Combining this equation and (53.7), we obtain

$$T = 2\pi/\Delta\omega = 2nL/c, \tag{53.18}$$

which means that T is the duration of a cycle. Thus, the situation in which the phases of various modes are equal at a certain instant of time is repeated after a cycle. The matching of phases of different modes is called mode synchronization. The instant when the amplitudes of modes are added in the same phase corresponds to a sharp increase in the intensity of laser radiation and a pulse is observed at this moment of time (Fig. 285). Pulses follow one another with a frequency $1/T = c/(2nL)$.

PULSE DURATION. This time is determined by the vanishing of the resultant amplitude. Let us denote by ΔT the pulse duration and by N the number of modes participating in

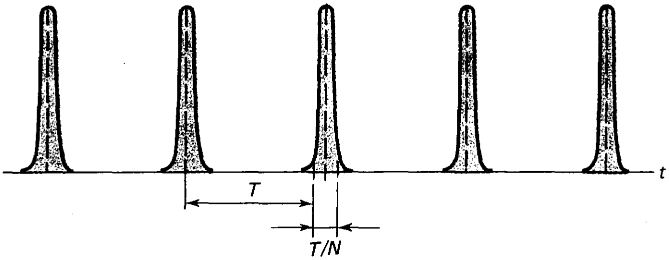


Fig. 285. Picosecond pulses in mode synchronization regime.

the process of synchronization. The condition under which the amplitude turns to zero is given, in analogy with (53.17), by

$$N\Delta\omega\Delta T = 2\pi, \qquad (53.19)$$

where

$$\Delta T = \frac{2\pi}{N\Delta\omega} = \left(\frac{2Ln}{c}\right)\frac{1}{N} = \frac{T}{N}. \qquad (53.20)$$

The larger the number of synchronized modes, the shorter the pulse duration. For example, we obtain $\Delta T \simeq 0.6 \times 10^{-12}$ s for $n \simeq 1$, $L = 10^{-1}$ m and $N = 10^3$. The mode synchronization method can be used to obtain the shortest possible durations of pulses which can be applied quite effectively for studying high-speed processes.

The output power for a pulse can be evaluated from the law of energy conservation. Due to mode synchronization, nearly all the energy emitted over an interval of time $T = 2Ln/c$ is released over a pulse of duration $\Delta T = T/N$. This means that the output power of the pulse increases by a factor of $T/\Delta T$ in comparison to the mean power.

ATTAINMENT OF MODE SYNCHRONIZATION. Mode synchronization may occur spontaneously. In this case, a small number of modes is usually synchronized and hence the pulse duration does not differ much from the period of their repetition (or the cycle duration). Different methods are employed for reducing the pulse duration, including Q-modulation with a period $2Ln/c$ for a cycle. This involves nonlinear effects which play a very significant role in laser physics. If nonlinearity is taken into account, different modes are interconnected and can no longer be treated as independent. Mode synchronization is not a completely random process any more, and it becomes possible to exert influence on it. As an example, let us consider lasing in the case when the resonator contains a filter whose transmission coefficient increases with the intensity of the light passing through it. The filter acts practically without any time lag. If a random synchronization leads to an increase in the intensity of a small

?

Which characteristics of laser radiation are determined by axial and side modes?
What factors determine the linewidth of laser radiation?
What is the theoretical limit of the relative radiation linewidth?
What relative linewidths of laser radiation have been attained at present?
How is mode synchronization achieved in lasers?

number of modes, these synchronized modes will be least attenuated upon passing through the filter and will prevail over other modes which pass through the filter at other instants of time with a smaller field strength and are therefore strongly attenuated. During subsequent passages through the filters, synchronized modes create conditions of relative amplification of other modes which turned out to be randomly synchronized with the existing modes. Thus, only synchronized modes or modes which happen to fall in the mode synchronization process by chance are amplified. This results in an increase in the number of synchronized modes, a decrease in the pulse duration, and a corresponding increase in the pulse intensity. Pulses with a duration much smaller than 10^{-12} s have been obtained at present.

LASER SPECKLES. The surface of most objects illuminated by a laser beam appears to be granular. The granules are randomly distributed over the surface and can be photographed. If the camera is focussed at a point above the surface or below it, the image of the granules is still formed on the film. If in the course of visual inspection of the granules an attempt is made to accommodate the eye at points not reaching the surface or lying beyond it, the sensory perception about the existence of the granules persists. If the observer is in motion while looking at the surface, some of the granules disappear while some new ones appear and it seems as though the granules twinkle and move relative to the surface. Such granules are termed speckles.

The emergence of speckles is due to a high degree of coherence of laser radiation. Most surfaces which have not been specially ground and polished as optically flat surfaces have random inhomogeneities whose height exceeds the wavelength of the laser radiation diffusely reflected by the surface with a very high degree of coherence. Waves with a constant phase difference propagate from different points of the surface. When these waves fall on the retina or a photographic film, an interference pattern in the form of alternating dark and bright speckles is observed.

If the surface is optically flat to within a fraction of a wavelength, the laser radiation is reflected without the introduction of any phase difference between the rays of the beam, and no speckles are observed. Speckles are also not observed in the case of diffuse reflection from the surface of a liquid where the inhomogeneities keep on varying with time, thus causing a variation of the path difference between the rays scattered from different points on the surface. This leads to a levelling out of the interference pattern with time.

From the physical point of view, the origin of speckles is associated with the noise effect. This effect has a number of useful applications. It can be used for measuring slight displacements of a solid surface, say, in a direction parallel to the surface. For this purpose, two consecutive photographs of the speckles are taken. If the surface remained immobile in the interval between the two instants, the speckles in the two photographs are superimposed on one another. In the case of a displacement absolutely two identical patterns of speckles, which are displaced relative to each other, are observed on the photographic plate. Knowing the relative positions of the photographic camera and the surface at the instants when the photographs were taken, we can easily calculate the displacement of the surface from the relative displacement of the speckle patterns on the photographic plate. In the same way, we can also determine the rotation of a surface about a certain axis. Speckles are also put to other applications.

Speckles are formed not only as a result of a diffuse scattering by a surface, but also when light passes through a scattering object, since the phase difference between the waves arriving at a given point from two different scatterers is always the same. The speckles formed in this way also have various applications.

SEC. 54. CHARACTERISTICS OF SOME LASERS

The characteristic features of the most typical lasers are briefly described.

VARIOUS TYPES OF LASERS. At present, there exists a large variety of lasers which differ in the choice of active medium, output power, operating conditions and other characteristics. Without going into a detailed description of all these characteristics, we shall briefly describe the lasers which offer a fairly detailed account of the characteristics of the basic types of laser (operating conditions, pumping techniques, etc.).

RUBY LASER. The first laser constructed in 1960 was the ruby laser.

The working medium in such a laser is a crystal of aluminium oxide Al_2O_3 (corundum) containing Cr_2O_3 as an impurity introduced during crystal growth. The red colour of the ruby crystal is due to the positive Cr^{3+} ion. The Cr^{3+} ion replaces the Al^{3+} ion in the lattice of Al_2O_3 crystal. This results in the formation of two absorption bands in the crystal, one of which is in the green region of the visible spectrum, and the other in the blue one (Fig. 286). The richness of the red colour of a ruby crystal depends on the concentration of Cr^{3+} ions: the higher the concentration, the richer the colour. In dark-red ruby, the concentration of Cr^{3+} ions is as high as 1%.

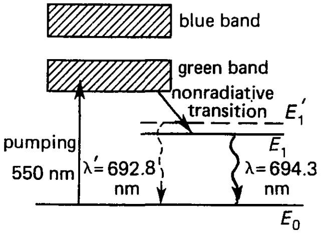


Fig. 286. Energy level diagram for ruby.

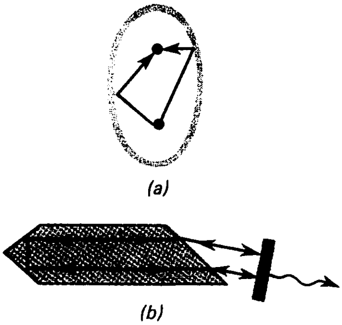


Fig. 287. Pumping (*a*) and path of rays in a ruby crystal (*b*).

In addition to the blue and green absorption band, a ruby crystal has two energy levels E_1 and E_1' corresponding to it. A transition from these levels to the ground state is accompanied by the emission of light with wavelengths 694.3 and 692.8 nm respectively. At room temperature, the width of these spectral lines is about 0.4 nm. The probability of induced transitions is higher for the wavelength of 694.3 nm than for 692.8 nm in view of the fact that $B \propto \omega^{-3}$. Hence it is more convenient to use the green line. However, it is possible to use the line 692.8 nm also with the help of special mirrors which have a higher reflectivity for radiation with $\lambda = 692.8$ nm than for radiation with $\lambda = 694.3$ nm.

The energy level diagram for ruby is shown in Fig. 286. When white light falls on ruby, the blue and the green regions of the spectrum are absorbed, while the red region is scattered. Optical pumping is carried out with the help of a xenon lamp emitting high-intensity pulses of light when a current pulse passing through it heats the gas to a temperature of several thousand kelvins. A continuous pumping is not possible because the lamp cannot operate continuously at such a high temperature. The radiation emitted by the lamp is similar in nature to the blackbody radiation, and is absorbed by Cr^{3+} ions which consequently go over to the energy levels in the absorption bands. As a result of nonradiative transition, however, Cr^{3+} ions rapidly go over to levels E_1 and E_1' (Fig. 286). The excess energy is transferred to the lattice, i. e. is transformed into the lattice vibration energy or, in other words, the phonon energy. Levels E_1 and E_1' are metastable. The lifetime corresponding to level E_1 is equal to 4.3 ms. Pulsed pumping leads to an accumulation of excited atoms on levels E_1 and E_1', resulting in a considerable population inversion with respect to level E_0.

The ruby crystal is grown in cylindrical form. Crystals having a length of about 5 cm and a diameter of about 1 cm are usually employed in lasers. The xenon lamp and the ruby crystal are placed in an elliptic cavity with highly reflecting inner surface (Fig. 287*a*). To ensure that the entire radiation from the xenon lamp falls on the ruby, the crystal and the lamp (which also has a cylindrical shape) are placed at the foci of the elliptic cross section of the cavity parallel to its generatrix. This ensures that the density of the radiation incident on ruby is the same as that of the radiation at the pumping source.

One of the faces of the ruby crystal is cut in such a way (Fig. 287*b*) that the rays incident on it are totally reflected backwards. Such a cut replaces one of the mirrors in the laser. The second face of the crystal is cut at the Brewster angle and allows the emergence of the ray from the ruby crystal without reflection and with an appropriate linear polarization. The second mirror of the resonator is placed in the path of this ray.

Thus, the radiation emitted by a ruby laser is linearly polarized.

HELIUM-NEON LASER. The active medium in such a laser is a gaseous mixture of helium and neon. Lasing is caused by transitions between neon energy levels, and helium plays the role of an intermediary through which energy is transferred to neon atoms for creating a population inversion.

In principle, neon can generate laser radiation as a result of more than 130 different transitions. However, the lines corresponding to the wavelengths 632.8 nm, 1.15 μm and 3.39 μm are the ones with the highest intensity (Fig. 288). The wavelength of 632.8 nm is in the visible part of the spectrum, while other wavelengths correspond to the infrared region.

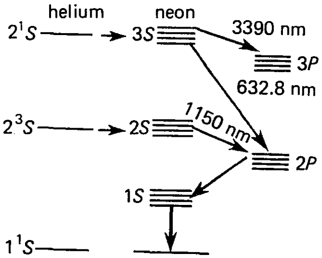


Fig. 288. Energy level diagram of helium and neon.

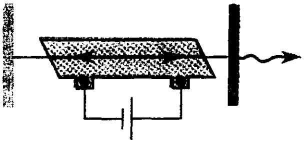


Fig. 289. Schematic diagram of a helium-neon laser.

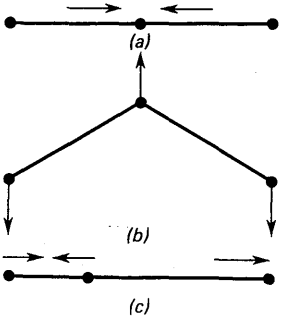


Fig. 290. Oscillation modes of a CO_2 molecule (*a*) symmetric extension, (*b*) bending, (*c*) non-symmetric extension.

When current is passed through a gaseous mixture of helium and neon, helium atoms are excited as a result of electron collisions to the levels 2^3S and 2^2S which are metastable since a transition from these levels to the ground state is forbidden by the quantum-mechanical selection rules. With the passage of current, atoms are accumulated on these levels. When an excited helium atom collides with an unexcited neon atom, the excitation energy is transferred to the latter. This transition takes place quite effectively on account of an ideal coincidence in the energies corresponding to these levels. Hence a population inversion on $3S$ and $2S$ levels is created in neon with respect to the levels $3P$ and $2P$, thus making laser action possible. The laser may function in the continuous mode. A typical diagram of a He-Ne laser is shown in Fig. 289. The ends of the laser tube are closed by an appropriate transparent material in such a way that the axial modes are incident on it at the Brewster angle. This ensures complete transmission of light of one polarization, and a complete elimination of the other polarization from the beam. The radiation from a helium-neon laser is linearly polarized. Usually, the pressure of helium and neon in the chamber is 332 Pa and 66 Pa respectively. A constant voltage of about 4 kV is applied across the tube. One of the mirrors has a reflectivity of about 0.999 while the other, through which the laser radiation passes, has a reflection coefficient of about 0.990. Multilayered dielectrics (see Sec. 29) are used as mirrors, since the lasing threshold cannot be attained for a lower reflectivity.

CLOSED-VOLUME CO_2-LASER. Like other molecules, the CO_2 molecules have a band spectrum formed by rotational and vibrational energy levels. The CO_2 molecule is a linear one

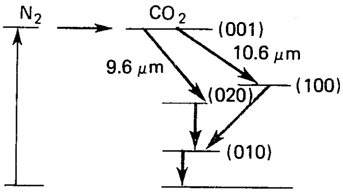


Fig. 291. Energy level diagram of a CO_2 laser.

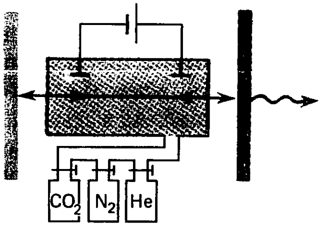


Fig. 292. Schematic diagram of a closed-system CO_2 laser.

and has a centre of symmetry. It has three fundamental vibrational modes (Fig. 290). The energy of the quanta of fundamental vibrational modes is given by (a) $1/\lambda_1 = 1337\ \text{cm}^{-1}$, (b) $1/\lambda_2 = 667\ \text{cm}^{-1}$ and (c) $1/\lambda_3 = 2349\ \text{cm}^{-1}$. There may be one or more quanta corresponding to each vibrational mode. The vibrational states of a molecule are denoted by the number of quanta in the corresponding fundamental vibrational mode. For example, (010) means that the symmetric and antisymmetric vibrations are not excited while the bending mode contains only one quantum. The transition used in a CO_2-laser gives radiation with a wavelength 10.6 μm, which lies in the infrared region of the spectrum. With the help of the vibrational levels, the emission frequency can be varied over an interval between 9.2 and 10.8 μm. Energy is transferred to CO_2 molecules by nitrogen molecules which are themselves excited as a result of collisions with electrons when a current is passed through the mixture. The energy level diagram for such a laser is shown in Fig. 291.

The excited state of an N_2 molecule is metastable and is separated by $2318\ \text{cm}^{-1}$ from the ground state, which means that it is very close to the energy level (001) of a CO_2 molecule. In view of the metastability of the excited state of N_2, the number of excited molecules keeps on increasing as a current is passed through the mixture. When nitrogen and carbon dioxide molecules collide, the excitation energy is resonantly transferred from N_2 to CO_2. This results in a population inversion between the levels (001), (100) and (020) of CO_2 molecules. Helium is usually introduced as an impurity to decrease the population density of the level (100) which has a long lifetime and a transition to this level deteriorates the lasing conditions. Under the usual conditions, the gaseous mixture in such a laser is composed of helium (1330 Pa), nitrogen (133 Pa) and carbon dioxide gas (133 Pa).

During the operation of a CO_2-laser, the CO_2 molecules dissociate into CO and O, as a result of which the active medium is depleted. In the next stage, CO is decomposed into C and O, and the carbon thus liberated is deposited on the electrodes and the tube walls. All this causes a deterioration of the laser operation. In order to eliminate the harmful effect of these factors, water vapours are introduced into the closed system. These vapours stimulate the reaction $CO + O \rightarrow CO_2$. Platinum, which serves as a catalyst for this reaction, is used as the material for the electrodes. To replete the stock of gases, the resonator is connected to additional sources of CO_2, N_2 and He. When required, these gases can be introduced into the closed volume of the laser to maintain the optimal proportion for a smooth running of the unit (Fig. 292). Such a closed-

!

Mode selection in lasers is attained by varying the Q-factor of the resonator for the vibrations corresponding to different modes. This allows us to control the number and characteristics of modes in laser radiation and vary the radiation frequency if mode selection is possible for a sufficiently wide frequency interval.

?

Why can a T-laser operate successfully if the working medium is under atmospheric pressure?
Which mechanism of creating population inversion is used in gasdynamic lasers?
How is mode selection realized in dye lasers?

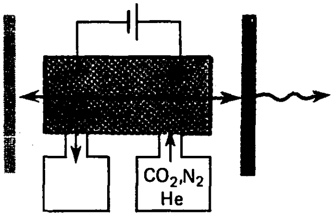


Fig. 293. Schematic diagram of a CW-mode CO_2 laser.

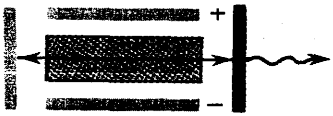


Fig. 294. Schematic diagram of a T-laser operating at atmospheric pressure.

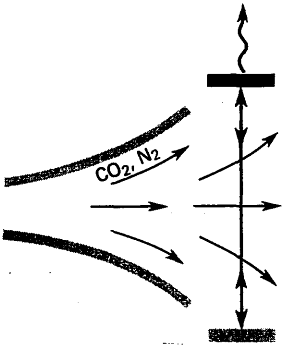


Fig. 295. Schematic diagram of a gasdynamic laser.

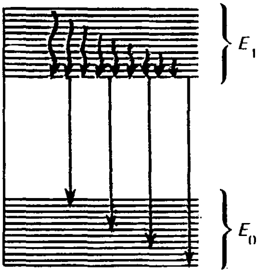


Fig. 296. Energy level diagram of a dye molecule.

system CO_2-laser is operative for several thousand hours.

CW-MODE CO_2-LASER. A modification of the CO_2-laser is the CW-mode laser in which a mixture of CO_2, N_2 and He gases is continuously pumped through the laser in the axial direction (Fig. 293). Such a laser is in a position to continuously generate coherent radiation whose power is more than 50 W per metre length of the active medium.

T-LASER. In CO_2-lasers used for many practical applications, the working mixture of the gases under atmospheric pressure is excited by a transverse electric field (T-laser, Fig. 294). Since the electrodes in such a laser are parallel to the axis of the resonator, a relatively small potential difference must be applied in order to produce a strong electric field in the resonator. This makes it possible to operate the laser in a pulsed mode under atmospheric pressure, when the concentration of CO_2 in the resonator is quite high. As a result, it is possible to develop high-power laser radiation. For a pulse whose duration is less than 1 μs, the output power is usually 10 MW or even higher. The pulse repetition frequency of such lasers is a few pulses per minute.

GASDYNAMIC LASERS. When a strongly heated mixture of CO_2 and N_2 (at temperatures between 1000 and 2000 K) is passed at a high velocity through an expansion nozzle, an intensive cooling takes place. In this case, the upper and lower energy levels are thermolyzed at different rates, and hence a population inversion is observed. This population inversion can be used to generate laser radiation if an optical resonator is mounted at the outlet from the nozzle (Fig. 295). Lasers based on this principle are called gasdynamic lasers and are capable of generating very high power of radiation in the CW-mode.

DYE LASERS. Dyes have very complex molecules with highly pronounced vibrational energy levels. The energy levels in a spectral band are almost continuous (Fig. 296). As a result of the intramolecular interaction, molecules undergo rapid (in about 10^{-11}–10^{-12} s) nonradiative transitions to the lowest energy level in each band. Hence, in a very short time following molecular excitation, all excited molecules are concentrated on the lowest level of the E_1-band. These molecules can then undergo a radiative transition to any energy level of the lower band (Fig. 296). Thus, radiation of practically any frequency can be emitted in the interval corresponding to the width of the lowest energy band. This means that if the dye molecules are used as the active medium for generation, laser radiation of any frequency can be obtained by an appropriate

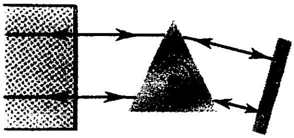


Fig. 297. Method for selection of lasing frequencies in a dye laser.

tuning of the resonator. Hence dyes are used for creating tunable lasers. Pumping in dye lasers is accomplished with the help of gas-discharge lamps or the radiation from some other type of lasers.

The separation of a lasing frequency is made possible by the fact that the lasing threshold is created only for a narrow frequency interval. For example, a mirror and a prism are arranged in such a way (Fig. 297) that after reflection by the mirror, only rays with a definite wavelength return to the medium on account of dispersion and different angles of refraction. Lasing is ensured only at these wavelengths. By rotating the prism, a continuous tuning of the laser radiation frequency can be ensured in dye lasers. Various types of materials are used in dye lasers. This makes it possible to obtain laser radiation not only in the entire visible range of the spectrum, but also over a considerable part of the infrared and ultraviolet region of the spectrum.

PROBLEMS

10.1. Two spherical cavities having large radii and small spherical apertures of diameter $d = 1$ cm each have totally reflecting outer surfaces. The planes of the circular apertures are parallel and their centres lie on a straight line perpendicular to these planes. The distance l between the apertures is 10 cm. A constant temperature T_0 of 2000 K is maintained in one of the cavities. Determine the steady-state temperature in the second cavity.

10.2. Knowing that the Sun emits radiation like blackbody at a temperature of 6000 K and that its diameter is seen from the Earth at an angle $\Delta\theta_0 = 32'$, determine the energy flux density of solar radiation on the Earth's orbit.

10.3. Find the temperature of the Sirius surface if the peak of the energy flux density of its radiation lies at the wavelength $\lambda_{\max} = 0.29$ μm.

10.4. A blackbody with a surface area of 10 m^2 has a radiation power of 1 MW. Find the temperature of this blackbody.

10.5. The radiation power of a blackbody is 100 MW, and the maximum radiation density is observed at a wavelength of 0.6 μm. Find the surface area of the blackbody.

10.6. A Lambert surface heated to 2900 K emits 200 W radiation from each square centimetre of its surface. Find the ratio of the radiant emittance of this surface to the radiant emittance of a blackbody.

10.7. A laser radiation pulse has an energy flux of 10^{11} W/cm^2. Find the peak value of the electric field strength in the wave.

10.8. What temperature of a blackbody corresponds to the radiation of a ruby laser with $\lambda = 694.3$ nm if the spectral density at this wavelength is $w_\omega = 10 \text{ J}\cdot\text{s}\cdot\text{m}^{-3}$?

ANSWERS

10.1. $T_0\pi^{1/4}[d/(2l)]^{1/2} = 595$ K. 10.2. 1.4 kW/m^2. 10.3. 10^4 K.
10.4. 1150 K. 10.5. 3.2 m^2. 10.6. 0.5. 10.7. 6.13×10^8 W/m.
10.8. 9.3×10^{16} K.

Chapter **11**

Nonlinear Phenomena in Optics

Basic idea:
For waves with quite high electric field strength, the dependence of polarization on electric field is nonlinear.

SEC. 55. NONLINEAR POLARIZATION

The relation between the polarization and the electric field strength is described by taking nonlinear terms into consideration.

LINEAR POLARIZATION. The theory of linear polarization, which established the frequency dependence of the refractive index, was described in Sec. 15. The polarization is connected with the electric field strength through the relation

$$\mathrm{P} = \varepsilon_0 \varkappa \mathrm{E}, \tag{55.1}$$

where $\varkappa$ is the dielectric susceptibility. A comparison of (55.1) and (15.13) leads to the following expression for the complex dielectric susceptibility:

$$\varkappa = \frac{e^2 N}{m\varepsilon_0} \frac{1}{\omega_0^2 - \omega^2 + i\gamma\omega}. \tag{55.2}$$

NONLINEAR POLARIZATION. The force $f(x) = -m\omega_0^2 x$, which holds an electron near its equilibrium position [see (9.32)], obeys Hooke's law only for values of x that are not too large. A departure from Hooke's law is observed for large x, and the oscillations become nonlinear.

In the general case, we can present the function $f(x)$ in the form of a Taylor series:

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \dots. \tag{55.3}$$

Taking into account the nonlinearity, we can replace (9.32) by the following expression

$$m\ddot{x} + m\gamma\dot{x} = eE + f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \dots, \tag{55.4}$$

which describes the motion of an electron in the field of a light wave. Since the point $x = 0$ corresponds to the equilibrium position, $f(0) = 0$. The force f is always directed towards the equilibrium position, and hence $f'(0) < 0$. Putting $f'(0) = -m\omega_0^2$, we can write (55.4) in the form

$$m\ddot{x} + m\gamma\dot{x} + m\omega_0^2 x = eE + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \dots . \tag{55.5}$$

Neglecting quadratic, cubic, and higher-order terms in x in (55.5), we arrive at Eq. (9.32) for a linear harmonic oscillator. When these terms are not neglected, the oscillator is called anharmonic, and its oscillations are called anharmonic oscillations. Obviously, the dependence $x(E)$, which is given by formula (9.36) for the linear case, becomes complicated for the anharmonic oscillations and is no longer linear. As a result, the polarization [see (15.13)], which can be conveniently written in the form

$$P = N|e|x, \tag{55.6}$$

is no longer a linear function of E.

QUADRATIC NONLINEARITY. If $f''(0) \neq 0$, the nonlinearity of oscillations is manifested primarily due to the quadratic term in x. Equation (55.5) can be written in the form

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = (e/m)E + \xi x^2, \tag{55.7}$$

where $\xi = f''(0)/(2m)$. Since the quantity ξx^2 is assumed to be small, Eq. (55.7) should be solved by the method of perturbations. We present the required solution in the form of the series

$$x = x_0 + x_1 + x_2 + \dots, \tag{55.8}$$

where $x_1, x_2, \dots$ are small quantities of the order of $\xi, \xi^2, \dots$ in x_0. Substituting (55.8) into (55.7) and equating terms of the same order in ξ, we arrive at the equations

$$\ddot{x}_0 + \gamma\dot{x}_0 + \omega_0^2 x = (e/m)E, \tag{55.9}$$

$$\ddot{x}_1 + \gamma\dot{x}_1 + \omega_0^2 x_1 = \xi x_0^2, \tag{55.10}$$

. . . .

where the dots stand for equations in $x_2, x_3, \dots$ which are not considered for the present.

In the linear case (55.1), we considered only one wave of frequency ω, since the addition of another wave of a different frequency did not add anything to the polarization pattern: the polarization due to two waves is equal to the sum of polarizations due to each wave. The situation changes if the nonlinearity is taken into consideration. It can be seen from (55.9) that x_0 is expressed as a linear function of E, while it follows from (55.10) that x_1 depends on x_0^2 and hence on E^2. If E is expressed as the sum of electric field strengths with different frequencies, x_1 will depend on the pairwise products of these field strengths, i.e. will be a quadratic function of field strengths. In accordance with (55.6), the polarization is described by the formula

$$P = N|e|x_0 + N|e|x_1 = P_l + P_n^{(2)}, \tag{55.11}$$

where

$$P_l = N|e|x_0 = \varepsilon_0 \varkappa^{(1)} E \tag{55.12}$$

and

$$P_n^{(2)} = N|e|x_1 = \varepsilon_0 \varkappa^{(2)} E^2 \tag{55.13}$$

represent the linear and quadratic nonlinear polarizations respectively.

Solving the equations denoted by dots and following Eq. (55.10), we obtain for x_2 a solution that depends on E^3 and leads to polarization P_n^3 which is proportional to E^3, and so on. This result can be written in symbolic form as follows:

$$\mathrm{P} = \varepsilon_0 \varkappa^{(1)} \mathrm{E} + \varepsilon_0 \varkappa^{(2)} \mathrm{EE} + \varepsilon_0 \varkappa^{(3)} \mathrm{EEE} + \ldots . \tag{55.14}$$

Here, the polarization is expressed as the sum of terms depending on the electric field strength linearly, quadratically, cubically, and so on.

NONLINEAR SUSCEPTIBILITY. In order to determine the quadratic nonlinear polarization, we must solve Eqs. (55.9) and (55.10). We shall assume that E is a superposition of real harmonic fields with frequencies $\omega_1, \omega_2, \omega_3, \ldots$:

$$E = E(\omega_1)e^{i\omega_1 t} + E^*(\omega_1)e^{-i\omega_1 t} + E(\omega_2)e^{i\omega_2 t}$$
$$+ E^*(\omega_2)e^{-i\omega_2 t} + \ldots + E(\omega_n)e^{i\omega_n t} + E^*(\omega_n)e^{-i\omega_n t}, \tag{55.15}$$

where the complex-conjugate pairs of the same frequency describe the real fields corresponding to these frequencies. In order to make the notation quite compact in the course of computations, it is expedient to use special symbols. The complex number

$$E(\omega_m)e^{i\omega_m t} + E^*(\omega_m)e^{-i\omega_m t} \tag{55.16}$$

is real if $E^*(\omega_m) = E(-\omega_m)$ [see (8.35)]. Hence, using the notation

$$\omega_{-m} = -\omega_m, \tag{55.17}$$

we can present the sum (55.15) in the form

$$E = \sum_{m=-n}^{n} E(\omega_m)e^{i\omega_m t}. \tag{55.18}$$

Substituting this relation into (55.9) and considering that for each term in the sum (55.18) the solution has a form similar to

(15.12), we obtain

$$x_0 = \frac{e}{m} \sum_{m=-n}^{n} \frac{E(\omega_m)\,\mathrm{e}^{i\omega_m t}}{\omega_0^2 - \omega_m^2 + i\gamma\omega_m}. \tag{55.19}$$

This leads to the following expression for the linear polarization:

$$P_l = N|e|x_0 = \varepsilon_0 \sum_{m=-n}^{n} \varkappa^{(1)}(\omega_m)\,E(\omega_m)\,\mathrm{e}^{i\omega_m t}, \tag{55.20}$$

where

$$\varkappa^{(1)}(\omega_m) = \frac{e^2 N}{m\varepsilon_0} \frac{1}{\omega_0^2 - \omega_m^2 + i\gamma\omega_m} \tag{55.21}$$

is the linear dielectric susceptibility which coincides with (15.13c) for each frequency. Substituting (55.19) into (55.10), we arrive at the following equation for x_1:

$$\ddot{x}_1 + \gamma\dot{x}_1 + \omega_0^2 x_1 = \frac{\xi e^2}{m^2} \times \sum_{\substack{m=-n\\k=-n}}^{\substack{k=n\\m=n}} \frac{E(\omega_m)\,E(\omega_k)\,\mathrm{e}^{i(\omega_m+\omega_k)t}}{(\omega_0^2 - \omega_m^2 + i\gamma\omega_m)(\omega_0^2 - \omega_k^2 + i\gamma\omega_k)}. \tag{55.22}$$

!

Under the influence of radiation, the polarization of a medium varies. As a result, the medium itself becomes a source of radiation which is composed with the original radiation in accordance with the superposition principle. Nonlinear phenomena in strong electromagnetic waves appear as a result of the nonlinear effect of the electromagnetic field on the polarization of the medium, and not according to the superposition principle for the electromagnetic field.

This equation is solved in the same way as Eq. (55.9) for x_0:

$$x_1 = \frac{\xi e^2}{m^2} \sum_{m,k} \frac{E(\omega_m)\,E(\omega_k)\,\mathrm{e}^{(\omega_m+\omega_k)t}}{(\omega_0^2 - \omega_m^2 + i\gamma\omega_m)(\omega_0^2 - \omega_k^2 - i\gamma\omega_k)} \times \frac{1}{[\omega_0^2 - (\omega_m + \omega_k)^2 + i\gamma(\omega_m + \omega_k)]}. \tag{55.23}$$

Here the limits for summation over m and k are the same as before, and have not been included in the formula for the sake of simplicity of notation. Hence we obtain the expression for the nonlinear polarization:

$$P_n^{(2)} = N|e|x_1 = \varepsilon_0 \sum \varkappa^{(2)}(\omega_m, \omega_k)\,E(\omega_m) \times E(\omega_k)\,\mathrm{e}^{i(\omega_m+\omega_k)t}, \tag{55.24}$$

where

$$\varkappa^{(2)}(\omega_m, \omega_k) = \frac{\xi_m \varepsilon_0^2}{N^2 e^3} \times \varkappa^{(1)}(\omega_m)\,\varkappa^{(1)}(\omega_k)\,\varkappa^{(1)}(\omega_m + \omega_k) \tag{55.25}$$

is the second-order susceptibility, describing the quadratic nonlinear polarization.

COMBINATION FREQUENCIES. It can be seen from (55.24) that the quadratic polarization contains terms with any

?

Which physical factors are responsible for the nonlinearity of polarization?
According to what rule are the combination frequencies generated?

combination frequencies $\omega_m + \omega_k$. For example, the quadratic polarization due to two fields of frequencies ω_1 and ω_2 depends on frequencies $\omega_1 + \omega_2$, $|\omega_1 - \omega_2|$, $2\omega_1$ and $2\omega_2$.

While calculating polarization of higher orders, we obtain similar expressions. The polarization depends on a combination of the frequencies involved. For example, the cubic polarization depends on the combination frequencies $\omega_m + \omega_k + \omega_p$, where the subscripts m, k, and p can have all possible values.

SEC. 56. GENERATION OF HARMONICS

Harmonic generation and conditions for vector and spatial synchronism are considered.

LINEAR POLARIZATION WAVES. The electric field of a wave propagating in a medium causes polarization at points of the medium. This polarization propagates in space in the form of polarization waves. At each point in the medium, the varying polarization generates a secondary electromagnetic wave which is superimposed on the wave responsible for polarization. The resultant wave is itself a source of polarization which in turn generates an electromagnetic wave that causes polarization, and so on to infinity. In other words, the polarization wave and the electromagnetic wave are mutually generated by each other.

Let us derive the equation for an electromagnetic wave, taking this relation into account. In the absence of a source, i.e. if $j = 0$ and $\rho = 0$, Maxwell's equations have the form

$$\text{curl } \mathbf{H} = \partial \mathbf{D}/\partial t, \tag{56.1}$$

$$\text{curl } \mathbf{E} = -\partial \mathbf{B}/\partial t = -\mu_0 \partial \mathbf{H}/\partial t. \tag{56.2}$$

The medium is assumed to be nonmagnetic ($\mu = \mu_0$). Differentiating both sides of Eq. (56.1) with respect to t, we obtain

$$\text{curl } (\partial \mathbf{H}/\partial t) = \partial^2 \mathbf{D}/\partial t^2. \tag{56.3}$$

Substituting into this equation the quantity $\partial \mathbf{H}/\partial t$ from (56.2), we obtain the equation

$$\text{curl curl } \mathbf{E} = -\mu_0 \partial^2 \mathbf{D}/\partial t^2. \tag{56.4}$$

Taking into account the equation

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}, \tag{56.5}$$

we can write (56.4) in the form

$$\text{curl curl } \mathbf{E} + (1/c^2)\partial^2 \mathbf{E}/\partial t^2 + \mu_0 \partial^2 \mathbf{P}/\partial t^2 = 0, \tag{56.6}$$

where $c^2 = 1/(\varepsilon_0 \mu_0)$. This equation describes the propagation of an electromagnetic wave taking into account its connection with polarization.

If only linear polarization is taken into account, Eq. (56.6) is linear. In this case, the mutual formation of an electromagnetic and a polarization wave by one another leads to a change in the velocity of propagation of the electromagnetic wave. Consequently, the polarization wave and the electromagnetic wave propagate with the same phase velocity and in the same phase. If, for example, an electromagnetic wave is propagating in the direction of positive values of z, the polarization wave also propagates in the same direction with the same phase velocity $c/n(\omega)$, where $n(\omega)$ is the frequency-dependent refractive index. These waves can be presented in the form

$$E = E_0 \cos \omega (t - zn/c), \tag{56.7}$$

$$P = \varepsilon_0 \varkappa E_0 \cos \omega (t - zn/c). \tag{56.8}$$

Due to a change in polarization, an electromagnetic wave is generated at each point of the medium. The amplitude of this wave is proportional to (56.8). Secondary waves, excited by polarization (56.8) at points z' and z'' and propagating in the direction of positive values of z, can be written in the form

$$E'(z, t) = A \cos \omega [t - z'n/c - (z - z')n/c], \tag{56.9a}$$

$$E''(z, t) = A \cos \omega [t - z''n/c - (z - z'')n/c]. \tag{56.9b}$$

Since

$$t - z'n/c - (z - z')n/c = t - zn/c,$$

$$t - z''n/c - (z - z'')n/c = t - zn/c,$$

we conclude that both waves (56.9a) and (56.9b), which originate at different points in the medium, arrive at point z in the same phase and are mutually augmented. This circumstance determines the possibility of propagation of an electromagnetic wave in a medium and its velocity. It can be stated that the secondary waves formed at various points in a linear medium are synchronous.

NONLINEAR POLARIZATION WAVES. Nonlinear quadratic polarization involves all possible combination frequencies of primary electromagnetic waves. Hence the secondary waves generated by it have the same combination frequencies and propagate with different velocities in accordance with the law of dispersion.

Superposition of waves having different frequencies is not of any interest since it does not lead to interference. Interference is possible only between waves of the same frequency, emitted at different points in the medium. If the waves are augmented as a result of interference, we can speak of the existence of a wave of the corresponding frequency in the medium. In other words, the generation of a new frequency can be interpreted as a nonlinear effect of wave propagation in the medium. If, however, there is no amplification, we do not observe any new frequency in the medium in spite of the fact that new frequencies are generated at each point of the medium. Let us consider the conditions under which the frequency of the generated waves is different from that of the primary electromagnetic wave. These conditions are called the conditions of spatial synchronism.

SPATIAL SYNCHRONISM CONDITION. Let us write explicitly the polarization waves generated by a quadratic polarization nonlinearity. Consider two primary electromagnetic waves with frequencies ω_1 and ω_2:

$$E_1 = E_{01} \cos(\omega_1 t - k_1 z), \tag{56.10}$$

$$E_2 = E_{02} \cos(\omega_2 t - k_2 z). \tag{56.11}$$

The quadratic polarization wave generated by them can be presented in the form

$$P = 2aE^2 = 2a(E_1 + E_2)^2 = 2a(E_{01}^2 \cos^2(\omega_1 t - k_1 z) + E_{02}^2 \cos^2(\omega_2 t - k_2 z) + 2E_{01} E_{02} \cos(\omega_1 t - k_1 z) \cos(\omega_2 t - k_2 z), \tag{56.12}$$

where the waves are described in the real form, and the coefficient $2a$ describes the second-order polarization. If required, the value of this coefficient can be calculated from (56.25). Considering the trigonometric relations

$$2\cos^2 \alpha = 1 + \cos 2\alpha, \quad 2\cos\alpha \cos\beta = \cos(\alpha + \beta) + \cos(\alpha - \beta),$$

we can present (56.12) in the form

$$P = P_0 + P_{2\omega_1} + P_{2\omega_2} + P_{\omega_1 + \omega_2} + P_{\omega_1 - \omega_2}, \tag{56.13}$$

where

$$P_0 = a(E_{01}^2 + E_{02}^2), \tag{56.14}$$

$$P_{2\omega_1} = aE_{01}^2 \cos[2(\omega_1 t - k_1 z)], \tag{56.15}$$

$$P_{2\omega_2} = aE_{02}^2 \cos[2(\omega_2 t - k_2 z)], \tag{56.16}$$

$$P_{\omega_1 + \omega_2} = 2aE_{01} E_{02} \cos[(\omega_1 + \omega_2) t - (k_1 + k_2) z], \tag{56.17a}$$

$$P_{\omega_1 - \omega_2} = 2aE_{01} E_{02} \cos[(\omega_1 - \omega_2) t - (k_1 - k_2) z]. \tag{56.17b}$$

Thus, in the case of a quadratic nonlinearity, two harmonic electromagnetic waves generate four polarization waves with frequencies $2\omega_1$, $2\omega_2$, $\omega_1 + \omega_2$ and $|\omega_1 - \omega_2|$, and a static polarization P_0. The emergence of static polarization is called optical detection by analogy with the detection of radio signals by rectification of the current, since the detected signal contains a constant component.

We shall derive the conditions of spatial synchronism by considering the example in which the frequency ω_1 is doubled. In this case, only one wave with frequency ω_1 may propagate in the medium, and the other wave with frequency ω_2 may be absent. The wave (56.15) can be presented in the form

$$P_{2\omega_1}(z, t) = aE_{01}^2 \cos 2\omega_1 [t - zn(\omega_1)/c], \tag{56.18}$$

where $n(\omega_1)/c = k_1/\omega_1$, $n(\omega_1)$ being the refractive index for the wave with frequency ω_1. By analogy with (56.9a) and (56.9b), the electromagnetic waves generated by the polarization waves at points z' and z'' can be described by the formulas

$$E'_{2\omega_1}(z, t) = A \cos \{2\omega_1 [t - z'n(\omega_1)/c - (z - z')n(2\omega_1)/c]\}, \tag{56.19a}$$

$$E''_{2\omega_1}(z, t) = A \cos \{2\omega_1 [t - z''n(\omega_1)/c - (z - z'')n(2\omega_1)/c]\}, \tag{56.19b}$$

where we have considered that after their formation at points z' and z'', these waves propagate with a velocity $c/[n(2\omega_1)]$ which is different from the velocity of propagation $c/[n(\omega_1)]$ of the polarization wave. Considering that

$$z'n(\omega_1)/c + (z - z')n(2\omega_1)/c = z'\Delta n/c + zn(2\omega_1)/c,$$
$$z''n(\omega_1)/c + (z - z'')n(2\omega_1)/c = z''\Delta n/c + zn(2\omega_1)/c,$$

where

$$\Delta n = n(\omega_1) - n(2\omega_1), \tag{56.20}$$

we can present (56.19a) and (56.19b) in the form

$$E'_{2\omega_1}(z, t) = A \cos \{2\omega_1 [t - zn(2\omega_1)/c - z'\Delta n/c]\}, \tag{56.21a}$$

$$E''_{2\omega_1}(z, t) = A \cos \{2\omega_1 [t - zn(2\omega_1)/c - z''\Delta n/c]\}. \tag{56.21b}$$

A comparison of these equations shows that the secondary waves arrive at point z in the same phase and augment each other only if

$$\Delta n = 0. \tag{56.22}$$

Taking (56.20) into account, we can write this condition in the form

$$n(\omega_1) = n(2\omega_1). \qquad (56.23)$$

This is the condition for spatial synchronism in the case of frequency doubling.

The second harmonic generation was first realized in 1961 when the frequency of the ruby laser radiation was doubled in a nonlinear crystal.

COHERENCE LENGTH. The phase difference at point z between the waves described by formulas (56.21a) and (56.21b) is

$$\Delta\varphi = 2\omega_1 \Delta n(z'' - z')/c = 2\omega_1 \Delta n l/c, \qquad (56.24)$$

where

$$l = z'' - z' \qquad (56.25)$$

is the distance between points z' and z'' where these waves are generated. Obviously, for $\Delta\varphi = 0$, the absolute value of the amplitude of the wave resulting from interference attains its peak equal to the sum of the absolute values of the amplitudes of the interfering waves. With increasing $\Delta\varphi$, the absolute value of the amplitude of the resultant wave decreases and becomes equal to zero for $\Delta\varphi = \pi$. Hence the secondary waves generated over the distance l_c, which satisfies the condition

$$2\omega_1 \Delta n l_c/c = \pi \qquad (56.26)$$

in accordance with (56.24), superimpose to form a wave with a nonzero amplitude. It follows from (56.26) that

$$l_c = \lambda/(4\Delta n), \qquad (56.27)$$

where $\omega_1/c = 2\pi/\lambda$ and l_c is the coherence length (or the characteristic distance over which the phase difference between secondary electromagnetic waves changes by less than π).

It can be clearly seen from (56.24) that the amplitude of the resultant secondary electromagnetic wave propagating in a medium varies periodically. We shall now determine the law of this periodic variation. If a primary wave enters the medium at point $z = 0$, the electric field strength of the resultant secondary wave at point z in the medium will be the sum of the electric field strengths of the secondary waves generated over the distance between 0 and z. In accordance with (56.21a), the electric field due to the radiators between z' and $z' + dz'$ is given by the expression

$$dE_{2\omega_1}(z, t) = A \cos [2\omega_1 (t - zn(2\omega_1)/c - z'\Delta n/c]\} dz'. \qquad (56.28)$$

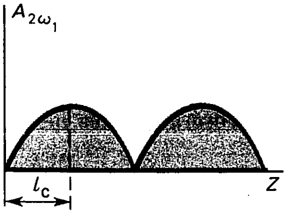


Fig. 298. Variation of the second harmonic amplitude with distance.

Consequently, the field strength of the secondary electromagnetic wave at point z is

$$E_{2\omega_1}(z, t) = A\int_0^z \cos\{2\omega_1[t - zn(2\omega_1)/c - z'\Delta n/c]\}\,\mathrm{d}z'. \quad (56.29)$$

Integrating, we obtain

$$E_{2\omega_1}(z, t) = \frac{Ac}{2\omega_1\Delta n}(\sin 2\omega_1[t - zn(2\omega_1)/c])$$
$$- \sin 2\omega_1[t - zn(2\omega_1)/c]), \quad (56.30)$$

where, in accordance with (56.20), $n(\omega_1) = n(2\omega_1) + \Delta n$. Using the trigonometric formula for the sine difference, we obtain

$$E_{2\omega_1} = A\frac{\sin(\omega_1\Delta nz/c)}{\omega_1\Delta n/c}\cos 2\omega_1$$
$$\times\left[t - \frac{n(2\omega_1) + n(\omega_1)}{2c}z\right]. \quad (56.31)$$

This formula shows that the secondary electromagnetic wave has double frequency $2\omega_1$, and that the resultant wave has a varying amplitude. The variation of the amplitude of the wave (Fig. 298) propagating through the medium is given by the function

$$A_{2\omega_1} = A\frac{\sin(\omega_1\Delta nz/c)}{\omega_1\Delta n/c}. \quad (56.32)$$

After the wave has entered the medium, its amplitude increases over a distance Δz if the following condition is satisfied:

$$\omega_1\Delta n\Delta z/c \leqslant \pi/2. \quad (56.33)$$

This is so because the value of the sine in (56.32) increases from 0 to 1. On the basis of (56.33), we can write the following expression for the distance over which the amplitude attains its maximum value:

$$\Delta z = \pi c/(2\omega_1\Delta n) = \lambda/(4\Delta n) = l_c. \quad (56.34)$$

This distance is equal to the coherence length (56.27). Hence, if spatial synchronism condition (55.27) cannot be satisfied for the generation of doubled frequency, we can try to generate this frequency by passing a high-power radiation through a plate whose thickness is equal to the coherence length. It has been shown experimentally that doubled frequency is observed at the exit from the plate for a high-power radiation. The intensity of doubled frequency generation is very low since the nonlinearity of the medium is generally quite small. To increase the radiation intensity, it is necessary to increase the coherence length by decreasing Δn. In the limit ($\Delta n \to 0$), the coherence length $l_c \to \infty$, which indicates the observance of condition (56.23) of spatial synchronism. The value of Δn depends on the properties of the medium and cannot be decreased arbitrarily. However, we can select conditions for anisotropic media under which $\Delta n = 0$. This is attained by using the difference between the refractive indices of ordinary and extraordinary rays.

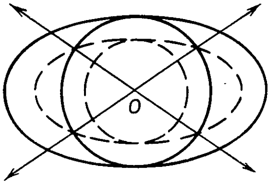


Fig. 299. Determining the directions of spatial synchronization in a uniaxial crystal.

Formula (56.32) describes the variation of the amplitude of the double-frequency wave shown in Fig. 298. The energy flux density of the wave is proportional to the square of this amplitude. Hence the process under consideration involves the transformation of the energy of a wave of frequency ω into the energy of a wave of frequency 2ω when the amplitude of the double-frequency wave increases and an inverse energy transformation when the amplitude of this wave decreases. In other words, this is a process of energy exchange between two waves having frequencies ω and 2ω, accomplished through the polarization of the medium.

ATTAINMENT OF SPATIAL SYNCHRONISM. In uniaxial crystals, ordinary and extraordinary rays have different velocities (see Sec. 41). The velocity of the ordinary ray does not depend on the direction, while that of the extraordinary ray depends on the angle between the ray and the optical axis. In positive crystals, the velocity of the ordinary ray is greater than that of the extraordinary ray, while the situation is reversed in negative crystals. Figure 299 shows the radiant surface cross sections of a negative uniaxial crystal by a plane passing through its optical axis. Since the frequencies corresponding to the visible, near ultraviolet and infrared regions lie in the normal dispersion region, the solid lines show the cross sections for the frequency ω, and the dashed lines correspond to the frequency 2ω. It can be seen that in the directions marked by two-pointed arrows, the velocities of the ordinary ray (frequency ω) and extraordinary ray (frequency 2ω) are equal, i.e. the corresponding refractive indices are equal:

$$n_0(\omega) = n_e(2\omega). \qquad (56.35)$$

This equality coincides with the spatial synchronism condition (56.23). However, it does not directly lead to the conclusion that all the arguments leading to formula (56.23) can be repeated and that the possibility of second harmonic generation can be reiterated for an extraordinary wave. As a matter of fact, the polarizations of ordinary and extraordinary waves are mutually perpendicular and hence the polarization wave associated with the ordinary wave cannot interact with the

!

The physical reason behind the frequency-doubling effect is the pumping of energy from one wave to another with the double frequency. The condition of spatial synchronism must be satisfied in order to ensure such an energy pumping in one direction for a long time. In the absence of spatial synchronism, energy is pumped into the wave with double frequency only for a short time after which the energy is pumped back from the wave with double frequency to the wave with the fundamental frequency, and the frequency-doubling effect is weakened.

extraordinary wave, and vice versa. If this is the case, we cannot use the arguments leading to condition (56.23) and state that (56.35) is the condition of spatial synchronism for the generation of the double-frequency extraordinary wave. However, such a statement is valid only for linear polarization. The nonlinearity is responsible for the interdependence of polarizations associated with ordinary and extraordinary waves. Hence all the arguments that led to formula (56.23) are valid for ordinary and extraordinary waves, and lead to Eq. (56.35) as the condition of spatial synchronism for the generation of an extraordinary wave of doubled frequency. Obviously, an ordinary wave of double frequency can be generated in a positive crystal.

It has been confirmed experimentally that the second harmonic generation in uniaxial crystals is highly effective when the direction of the rays coincides with that of spatial synchronism. The coefficient of transformation of energy of radiation with the original frequency into the energy of radiation with double frequency attains values of several tens percent. This phenomenon can be demonstrated quite effectively in the classroom. For this purpose, an invisible laser radiation of wavelength 1.06 μm is made to pass through a crystal whose orientation is varied. When the direction of spatial synchronism for the crystal coincides with the direction of the ray, green light (of wavelength 0.53 μm) is observed in the crystal in the form of a flash if the orientation of the crystal is continuously changed. If, however, the crystal is fixed in this position, the green light is emitted either continuously or in the form of pulses, depending on the nature of laser radiation incident on the crystal.

GENERATION OF HIGHER HARMONICS. The third-order nonlinear polarization contains terms with frequency 3ω, i.e. thrice that of the incident radiation. In complete analogy with the situation considered above, this leads to the generation of the third harmonic. The fourth and higher harmonics can be generated in a similar manner. This requires a perfection of the experimental techniques, but does not contain anything new in the physical sense.

VECTOR CONDITION FOR SPATIAL SYNCHRONISM. At first glance, it seems that the existence of the polarization waves (56.17a) and (15.17b) may lead to the generation of the waves with frequencies $\omega_1 + \omega_2$ and $|\omega_1 - \omega_2|$ in the same way as the generation of the doubled-frequency waves. This, however, is not true. As a matter of fact, in accordance with the quantum-mechanical interpretation, the monochromatic waves are photons with a definite frequency and momentum. Hence the creation of an electromagnetic wave of frequency $\omega_1 + \omega_2$ due to the polarization $P_{\omega_1+\omega_2}$ may be considered as a merger of photons with frequencies ω_1 and ω_2. Here the laws of conservation of energy and momentum should be obeyed simultaneously. The energy and momentum of the original photons can be written in the form

$$\varepsilon_1 = \hbar\omega_1, \quad p_1 = \hbar k_1 = h\omega_1 n(\omega_1)/c, \tag{56.36}$$

$$\varepsilon_2 = \hbar\omega_2, \quad p_2 = \hbar k_2 = \hbar\omega_2 n(\omega_2)/c. \tag{56.37}$$

Assuming that the electromagnetic wave due to polarization (56.17a) can also be represented by a photon, the energy and momentum of the latter can be expressed in the form

$$\varepsilon = \hbar(\omega_1 + \omega_2), \quad p = \hbar(\omega_1 + \omega_2)n(\omega_1 + \omega_2)/c. \tag{56.38}$$

It is obvious that in this case, the law of energy conservation

$$\varepsilon = \varepsilon_1 + \varepsilon_2 \tag{56.39}$$

is satisfied, while that of momentum conservation

$$p \neq p_1 + p_2 \tag{56.40}$$

is violated, since

$$\omega_1 n(\omega_1) + \omega_2 n(\omega_2) \neq (\omega_1 + \omega_2)n(\omega_1 + \omega_2). \tag{56.41}$$

In the case of harmonic doubling ($\omega_1 = \omega_2 = \omega$), inequality (56.41) is transformed into the equality $2\omega n(\omega) = 2\omega n(2\omega)$ if the condition $n(\omega) = n(2\omega)$ of spatial synchronism is satisfied. The law of momentum conservation is also obeyed in this case.

In order to apply the mechanism of double-frequency generation to the generation of sum and difference frequencies, we must consider electromagnetic waves propagating in an arbitrary direction in a medium. Instead of the waves (56.10) and (56.11) which propagate in the direction of the Z-axis, let us consider waves characterized by the wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$:

$$E_1 = E_{01}\cos(\omega_1 t - \mathbf{k}\cdot\mathbf{r}), \tag{56.42}$$

$$E_2 = E_{02}\cos(\omega_2 t - \mathbf{k}_2\cdot\mathbf{r}). \tag{56.43}$$

For the polarization, we obtain relations similar to (56.15)-(56.17) in which $k_1 z$ and $k_2 z$ have been replaced by $\mathbf{k}_1\cdot\mathbf{r}$ and $\mathbf{k}_2\cdot\mathbf{r}$ respectively. In particular, we obtain

$$P_{\omega_1+\omega_2} = 2aE_{01}E_{02}\cos[(\omega_1 + \omega_2)t - (\mathbf{k}_1 + \mathbf{k}_2)\cdot\mathbf{r}]. \tag{56.44}$$

The wave generated by this polarization can be considered as a plane monochromatic wave with wave vector $\mathbf{k}$ and frequency ω only if the laws of energy and momentum conservation are satisfied:

$$\omega = \omega_1 + \omega_2. \tag{56.45}$$

$$\mathbf{k} = \mathbf{k}_1 + \mathbf{k}_2. \tag{56.46}$$

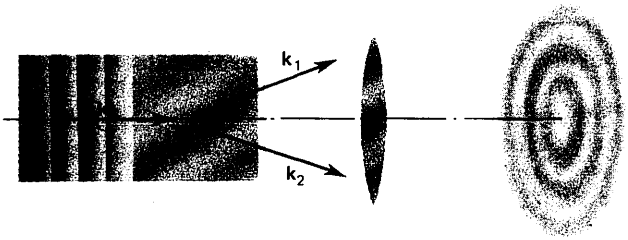


Fig. 300. Observing the spontaneous decay of a photon.

The last equality is called the condition of vector spatial synchronism. When applied to the case of frequency doubling of waves propagating in the same direction, as was explained in connection with (56.41), leads to condition (56.23) of spatial synchronism.

If there is dispersion, equality (56.46) cannot be satisfied [in view of (56.41)] for isotropic media even for collinear vectors k_1 and k_2, not to speak of noncollinear vectors. Hence we must look for the conditions under which the vector spatial synchronism is observed in anisotropic media. With the help of ordinary and extraordinary rays for certain angles between wave vectors, this condition can be satisfied. For example, in the potassium dihydrophosphate (KH_2PO_4) crystal, this condition can be obeyed in two ways:

$$k^e = k_1^{(o)} + k_2^{(o)}, \tag{56.47}$$

$$k = k_1^{(o)} + k_2^{(o)}. \tag{56.48}$$

GENERATION OF SUM AND DIFFERENCE FREQUENCIES. If the waves characterized by the vectors k_1 and k_2 satisfy condition (56.46) of vector spatial synchronism, the polarization wave (56.44) generates an electromagnetic wave of frequency $\omega_1 + \omega_2$ with the maximum intensity. But this does not mean that waves with frequency $\omega_1 + \omega_2$ will not be generated for any other ratio of vectors k_1 and k_2. Like in the case when the condition of spatial synchronism is not met for double-frequency generation, this generation occurs over small coherence lengths and is quite weak. The generation of the frequency difference $|\omega_1 - \omega_2|$ and other combinations of frequencies are analyzed in a similar manner.

SPONTANEOUS DECAY OF A PHOTON. If light passes through a crystal which is suitably oriented in such a way that condition (56.46) is satisfied, some of the photons of the incident radiation decay into two photons having lower frequencies ω_1 and ω_2, and moving at a certain angle to the direction of propagation of the initial photon (Fig. 300). This phenomenon is called the spontaneous decay of a photon and

?

What is the physical meaning of spatial synchronism? How is it achieved? What is the physical significance of the condition of vector spatial synchronism?
Describe the principle of parametric amplification of light and its application in the parametric generators of light.

can be observed experimentally. The pattern of photon decay is axially symmetric. The spot at the centre is formed by photons which pass through the crystal without undergoing a decay. This spot is surrounded by a coloured circle, the colouring changing from the centre to the periphery (the wavelength decreases with distance from the axis in complete agreement with the requirement of vector spatial synchronism).

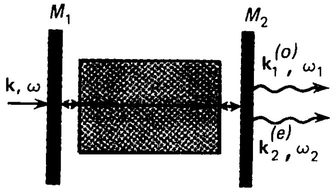


Fig. 301. Schematic diagram of a parametric generator of light.

PARAMETRIC AMPLIFICATION OF LIGHT. If three waves satisfying condition (56.46) propagate in a medium, an exchange of energy takes place between them. If one of the waves (say, with wave vector k) is much more powerful than the other two waves, the energy is transferred from this wave to the weaker waves which are consequently amplified. This phenomenon is called parametric amplification of light, since it can be considered as the modulation of the optical parameters of the medium by the wave k, leading to the amplification of waves k_1 and k_2.

PARAMETRIC GENERATORS OF LIGHT. These generators are based on the parametric amplification of light. A parametric generator is a laser whose cavity contains a nonlinear crystal cut in such a way that one of the vector conditions of spatial synchronism, viz. $k_1^{(o)} + k_2^{(o)} = k^{(e)}$ or $k_1^{(o)} + k_2^{(e)} = k^{(e)}$ (Fig. 301), is satisfied for the axial rays. The lasing frequencies of the generator are ω_1 and ω_2, and the reflectivities of the mirrors M_1 and M_2 are chosen in accordance with these frequencies. The generator is pumped by a beam of light with frequency ω, and the mirror M_1 should be quite transparent for this frequency.

If the orientation of the crystal or its parameters are changed (say, by varying the temperature, external electric field, etc.) the conditions of spatial synchronism change for axial rays, resulting in a change in the frequency of the laser radiation. This affords a possibility of a continuous variation of the generation frequency.

SEC. 57. SELF-FOCUSSING OF LIGHT IN A NONLINEAR MEDIUM

The basic laws of self-focussing and defocussing of light in a nonlinear medium are considered.

NONLINEAR CORRECTION TO THE REFRACTIVE INDEX. By definition, the following relation must be obeyed for isotropic media or crystals having a centre of symmetry:

$$P(-E) = -P(E). \qquad (57.1)$$

This means that all the terms containing an even number of multipliers E in (55.14) must vanish. Hence the first nonlinear

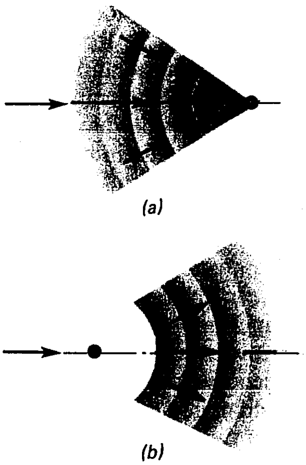


Fig. 302. Focussing and defocussing in a nonlinear medium: (*a*) $n_2 < 0$, $n_2 > 0$.

correction is given by the cubic term and the polarization assumes the form

$$\mathrm{P} = \varepsilon_0 \varkappa \mathrm{E} + a_3 E^2 \mathrm{E}. \tag{57.2}$$

This gives

$$\mathrm{D} = \varepsilon \mathrm{E} = \varepsilon_0 \mathrm{E} + \mathrm{P} = [\varepsilon_0 (1 + \varkappa) + a_3 E^2]\, \mathrm{E}. \tag{57.3}$$

The permittivity is then given by

$$\varepsilon = \varepsilon_0 (1 + \varkappa) + a_3 E^2. \tag{57.4}$$

The refractive index is defined by the relation

$$n = \sqrt{\varepsilon/\varepsilon_0} = \sqrt{1 + \varkappa + a_3 E^2/\varepsilon_0} \approx \sqrt{1 + \varkappa} \left(1 + \frac{a_3 E^2}{2\varepsilon_0 (1 + \varkappa)} \right), \tag{57.5}$$

since the nonlinear correction to the refractive index is much smaller than unity:

$$a_3 E^2/[\varepsilon_0 (1 + \varkappa)] \ll 1. \tag{57.6}$$

Thus,

$$n = n_0 + n_2 E_0^2. \tag{57.7}$$

Here, the square of the electric field strength has been averaged over a period of oscillations, and hence E_0 is the amplitude of the electric field strength of the wave; n_0 is the linear refractive index of light and $n_2 E_0^2$ describes the nonlinear correction to the refractive index.

SELF-FOCUSSING AND DEFOCUSSING OF A BEAM. The energy flux density over the cross section of a beam is not constant, and hence refractive index (57.7) also varies over the beam cross section. For the sake of definiteness, we shall assume that the beam has a circular cross section and that the intensity distribution is Gaussian (see Sec. 3). Depending on the nature of the nonlinearity, the sign of the correction to (57.7) may be positive or negative.

If the nonlinear correction is positive, the peripheral regions move faster than the central regions. As a result, the plane wave front becomes concave in the direction of propagation of the beam (Fig. 302*a*) and is focussed at the axis.

If the nonlinear correction is negative, the central part of the beam moves with a higher velocity than the peripheral regions. Consequently, the plane wave front becomes convex towards the direction of propagation (Fig. 302*b*) and is defocussed from the axis.

SELF-FOCUSSING LENGTH. Let us denote by a the radius of a beam and by E_0 the amplitude of the electric field strength at the axis. The amplitude at a distance a from the

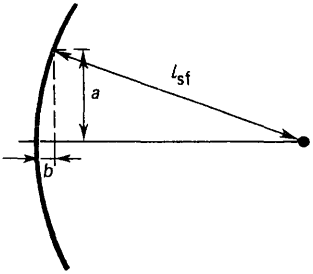


Fig. 303. To the calculation of the self-focussing length.

axis is assumed to be equal to zero (Fig. 303). By definition, the self-focussing length is the distance l_{sf} over which the beam is focussed at the axis in a nonlinear medium or, in other words, collapses to the axis.

Over the distance l_{sf}, a phase difference

$$\Delta\varphi = \omega\Delta n l_{sf}/c, \tag{57.8}$$

where

$$\Delta n = n_2 E_0^2, \tag{57.9}$$

appears between the axial ray and the extreme rays. It can be seen from Fig. 303 that

$$b = l_{sf} - \sqrt{l_{sf}^2 - a^2} \approx a^2/(2l_{sf}), \tag{57.10}$$

where $a/l_{sf} \ll 1$. For the bending of the wavefront shown in Fig. 303, the phase difference is

$$\Delta\varphi_1 = \omega n_0 b/c \approx \omega n_0 a^2/(2cl_{sf}), \tag{57.11}$$

where we have used expression (57.10) for b. The phase differences (57.8) and (57.11) must be equal:

$$\omega\Delta n l_{sf}/c = \omega n_0 a^2/(2cl_{sf}), \tag{57.12}$$

whence

$$l_{sf} = a\sqrt{n_0/(2\Delta n)} = (a/E_0)\sqrt{n_0/(2n_2)}. \tag{57.13}$$

The linear dependence of l_{sf} on a/E_0 is confirmed quite well in experiments. The effect becomes noticeable even for rather small values of $\Delta n/n$. From (57.13), we obtain

$$\Delta n/n = a^2/(2l_{sf}^2). \tag{57.14}$$

For example, when $\Delta n/n \simeq 0.5 \times 10^{-4}$, we obtain $a/l_{sf} \simeq 10^{-2}$, which is easily observable.

Let us express E_0 in terms of the radian energy flux density, assuming that the amplitude distribution of the waves in the beam has the form

$$E(r) = E_0\sqrt{1 - r^2/a^2}. \tag{57.15}$$

For the mean value of the energy flux density, we obtain the expression

$$\langle S\rangle = v\varepsilon E^2/2 = [\varepsilon c E_0^2/(2n_0)](1 - r^2/a^2)$$
$$= (\varepsilon_0 n_0 c E_0^2/2)(1 - r^2/a^2). \tag{57.16}$$

The energy flux of the beam is given by

$$P = \int_\sigma \langle S\rangle \, \mathrm{d}\sigma = \pi\varepsilon_0 n_0 c E_0^2 \int_0^a (1 - r^2/a^2) r \, \mathrm{d}r, \tag{57.17}$$

!

Physically, self-focussing and defocussing involve the deformation of a wave front due to a difference in their velocities of propagation on account of a change in the refractive index of the medium under the action of waves. In other words, the medium as if plays the role of an intermediary through which the waves act on themselves.

?

How can one determine whether a beam of light will be self-focussed or defocussed?
Which physical factors determine the threshold power of radiation?
What are the main reasons behind the nonlinearity of refractive index?

where σ is the cross-sectional area of the beam. Integrating, we obtain

$$P = \pi\varepsilon_0 n_0 c E_0^2 a^2/4 = \pi\varepsilon_0 n_0 c a^4/(8 n_2 l_{cf}^2). \tag{57.18}$$

For carbon bisulphide, for example, $n_0 = 1.62$, $n_2 = 0.22 \times 10^{-19}\ \mathrm{m^2/W^2}$, and a power of 10 MW is required for obtaining $l_{sf} = 10$ cm at $a = 1$ cm.

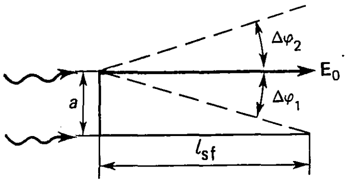


Fig. 304. To the calculation of the threshold energy flux of a beam for self-focussing.

THRESHOLD ENERGY FLUX. The energy flux required for the self-focussing of a beam decreases with the beam radius a. However, a decrease in the beam radius means an increase in the diffraction divergence which can be removed only by increasing the energy flow rate in the beam. The threshold energy flux is the minimum energy flow rate leading to a collapse of the beam. Its value can be determined from the following considerations.

On account of the nonlinearity, the beam collapses at a distance l_{sf}. This means that the peripheral ray of the beam (Fig. 304) deviates from the axis by an angle

$$\Delta\varphi_1 \approx a/l_{sf}. \tag{57.19}$$

As a result of diffraction, the peripheral ray is declined at an angle

$$\Delta\varphi_2 \approx \lambda/(a n_0) \tag{57.20}$$

from the beam axis. The threshold energy flux is determined from the condition

$$\Delta\varphi_1 = \Delta\varphi_2, \tag{57.21}$$

which assumes the following form if we take into account (57.19) and (57.20):

$$a/l_{sf} = \lambda/(a n_0). \tag{57.22}$$

Taking (57.13) into account, we obtain

$$a^2 = \lambda l_{sf}/n_0 = [\lambda a/(n_0 E_0)]\sqrt{n_0/(2n_2)}. \tag{57.23}$$

Taking the squares of both sides, we get

$$a^2 E_0^2 = \lambda^2/(2 n_0 n_2). \tag{57.24}$$

Substituting this expression into (57.18), we get

$$P_{thr} = \pi\varepsilon_0 c \lambda^2/(8 n_2). \tag{57.25}$$

For $P > P_{thr}$, the beam of light is focussed, while for $P < P_{thr}$, divergence is observed. The threshold energy flux is independent of the beam radius and decreases with decreasing wavelength. For example, when carbon bisulphide is exposed to ruby laser radiation with $\lambda = 694.3$ nm, we have

$$P_{thr} = 23\ \mathrm{kW}. \tag{57.26}$$

MAIN REASONS BEHIND NONLINEARITY OF THE REFRACTIVE INDEX.

1. It was explained during a discussion of the dispersion phenomena that the refractive index depends on absorption which, in turn, depends on the radiation density (see Sec. 51). Consequently, the refractive index depends on the square of the electric field strength of the wave.

2. As a result of electrostriction caused by the electric field of a wave, the density of a dielectric changes. The pressure at the surface of a volume element of the dielectric is given by $p \propto \rho (\partial \varepsilon/\partial \rho)$, while the change in the refractive index due to a change in the density of the dielectric is $\Delta n \propto p(\partial n/\partial \rho)$ and depends on the energy flux density.

3. The density of a medium, and hence its refractive index, change as a result of heating. The heating of the medium depends on the energy flux density of the wave. Since the energy flux density at the axis is higher than at the periphery, the central parts of the beam are heated more strongly, and their refractive index decreases. Consequently, heating results in a defocussing of the beam.

4. In general, molecules have different polarizations in different directions. In low external fields, molecules have random orientations and hence there is no anisotropy in the polarization properties of the medium. In strong fields, molecules are oriented in a definite manner with respect to the applied field and hence the polarization and refractive index become anisotropic, while the medium is transformed into a uniaxial crystal from optical point of view. Birefringence is observed in this case and the refractive index n of the extraordinary ray depends on the direction of propagation. The nonlinearity created due to this effect is orientational.

TIME LAG. Electrostrictive variation depends on the time required for establishing the density distribution in the medium. The velocity of an elastic wave in the medium is $v_3 \simeq 1.5 \times 10^3$ m/s, and hence the time lag τ is of the order $a/v_3 \simeq 10^{-6}$ s for $a \simeq 1$ mm.

The time taken by molecules for reorientation is of the order of 10^{-11}-10^{-12} s. Hence the orientation (Kerr) mechanism is the basic factor responsible for the emergence of nonlinearity for short laser pulses of duration shorter than 10^{-7} s. For longer pulses, we must also take into account other mechanisms that are responsible for the nonlinearity. The relative role of various mechanisms depends on the properties of the material and the characteristics of radiation pulses.

Appendix

SI Units Used in the Book

Quantity			Unit	
name	**dimensions**	**symbol**	**name of unit**	**symbol**
		Base units		
length	L	l	metre	m
mass	M	m	kilogram	kg
time	T	t	second	s
current	I	I	ampere	A
thermodynamic temperature	θ	T	kelvin	K
amount of substance	N	ν	mole	mole
luminous intensity	J	I	candela	cd
		Derived units		
phase velocity	LT^{-1}	v_{ph}, v	metre per second	m/s
group velocity	LT^{-1}	v_{g}, v	metre per second	m/s
velocity of light in vacuum	LT^{-1}	c	metre per second	m/s
acceleration	LT^{-2}	a	metre per second squared	m/s^2
force	LMT^{-2}	F	newton	N
light pressure	$\mathrm{L}^{-1}\mathrm{MT}^{-2}$	p	pascal	Pa
energy	$\mathrm{L}^2\mathrm{MT}^{-2}$	W, E	joule	J
volumetric energy density	$\mathrm{L}^{-1}\mathrm{MT}^{-2}$	w	joule per cubic metre	$\mathrm{J/m}^3$
power	$\mathrm{L}^2\mathrm{MT}^{-3}$	P	watt	W
energy flux density	MT^{-3}	S	watt per square metre	$\mathrm{W/m}^2$
momentum flux density	$\mathrm{L}^{-2}\mathrm{MT}^{-1}$	G	kilogram second per square metre	$\mathrm{kg\cdot s/m}^2$
area	L^2	S, σ	square metre	m^2
volume	L^3	V	cubic metre	m^3
cyclic frequency	T^{-1}	ω, Ω	reciprocal second	s^{-1}
frequency of a periodic process	T^{-1}	ν	hertz	Hz
wavelength	L	λ	metre	m
coherence time	T	τ_{c}	second	s
pulse duration	T	τ, Δt	second	s
coherence length	L	l_{c}	metre	m
coherence width	L	a_{c}	metre	m
wave number	L^{-1}	k	reciprocal metre	m^{-1}

oscillation phase	dimensionless	φ	radian	rad
plane angle	”	α, β, θ	radian	rad
solid angle	”	Ω	steradian	sr
oscillation period	T	T	second	s
damping factor	T^{-1}	γ	reciprocal second	s^{-1}
gain factor	L^{-1}	α	reciprocal metre	m^{-1}
Q-factor	dimensionless	Q		
refractive index	”	n		
reflectivity	”	ρ		
transmission coefficient	”	τ		
focal length	L	f	metre	m
phase difference	dimensionless	δ	radian	rad
optical path	L	Δ	metre	m
refractive index of ordinary ray	dimensionless	n_o		
refractive index of extraordinary ray	”	n_e		
volumetric charge density	$\mathrm{L}^{-3}\mathrm{TI}$	ρ	coulomb per cubic metre	$\mathrm{C/m}^3$
absolute permittivity	$\mathrm{L}^{-3}\mathrm{M}^{-1}\mathrm{T}^4\mathrm{I}^2$	ε	farad per metre	F/m
permittivity of vacuum	$\mathrm{L}^{-3}\mathrm{M}^{-1}\mathrm{T}^4\mathrm{I}^2$	ε_0	farad per metre	F/m
relative permittivity	dimensionless	ε_r		
electric field strength	$\mathrm{LMT}^{-3}\mathrm{I}^{-1}$	E	volt per metre	V/m
electric dipole moment	LTI	p	coulomb metre	$\mathrm{C \cdot m}$
polarization	$\mathrm{L}^{-2}\mathrm{TI}$	P	coulomb per square metre	$\mathrm{C/m}^2$
electric displacement	$\mathrm{L}^{-2}\mathrm{TI}$	D	coulomb per square metre	$\mathrm{C/m}^2$
magnetic induction	$\mathrm{MT}^{-2}\mathrm{I}^{-1}$	B	tesla	Tl
magnetic field strength	$\mathrm{L}^2\mathrm{MT}^{-2}\mathrm{I}^{-1}$	H	ampere per metre	A/m
absolute magnetic permittivity	$\mathrm{LMT}^{-2}\mathrm{I}^{-2}$	μ	henry per metre	H/m
permeability of vacuum	$\mathrm{LMT}^{-2}\mathrm{I}^{-2}$	μ_0	henry per metre	H/m
relative permeability	dimensionless	μ_r		
spectral density of radiation on cyclic frequency scale	$\mathrm{L}^{-1}\mathrm{MT}^{-1}$	w_ω	joule second per cubic metre	$\mathrm{J \cdot s/m}^3$
spectral density of radiation on wavelength scale	$\mathrm{L}^{-2}\mathrm{MT}^{-2}$	w_λ	joule per metre to fourth power	$\mathrm{J/m}^4$
luminous intensity	$\mathrm{L}^2\mathrm{MT}^{-3}$	I	watt per steradian	W/sr
radiant illuminance	MT^{-3}	E	watt per square metre	$\mathrm{W/m}^2$
radiance	MT^{-3}	L	watt per steradian square metre	$\mathrm{W\ sr}^{-1}\ \mathrm{m}^{-2}$
radiant emittance	MT^{-3}	M	watt per square metre	$\mathrm{W/m}^2$
luminous flux	J	Φ_V	lumen	lm
illuminance	$\mathrm{L}^{-2}\mathrm{J}$	E_V	lux	lx
luminance	$\mathrm{L}^{-2}\mathrm{J}$	L_V	candela per square metre	$\mathrm{cd/m}^2$
luminous emittance	$\mathrm{L}^{-2}\mathrm{J}$	M_V	lux	lx

Conclusion

This volume from the course of general physics* completes the four-volume treatise on classical physics. In short, classical physics can be defined as that branch of physics in which the role of quantum regularities is negligibly small. At first glance, it seems that this definition is devoid of any meaning since it describes what classical physics is not rather than what it actually is. However, such a view is deceptive since the discipline of physics is a single entity which is essentially of quantum nature. Hence the definition of classical physics as a branch of physics in which the role of quantum regularities is negligibly small is undoubtedly meaningful. Nevertheless, it is expedient to discuss the main properties of classical physics without referring to quantum laws.

The basic concepts of classical mechanics are those of a material body, a material point, the motion of a material point along a certain trajectory, and the forces which are supposed to be responsible for the regularities in the motion of material bodies and points. Although the modern concepts of classical physics are assumed to be formulated from Newton's times, some fundamental notions of classical physics date back to much earlier periods. These were developed gradually in human consciousness from ancient times as a result of practical day-to-day work. The practical activity of man also indicated that all material bodies have finite dimensions, occupy a certain volume and are arranged in a definite manner relative to one another. These general properties of material bodies created in our consciousness the concept of space, while the mathematical formulation of these properties was reflected in the form of a system of geometrical concepts and the relations between them. Practical activity of man also led to the conclusion that the material world surrounding him is in the process of continuous variation. The property of material processes to have a certain duration, to follow a certain sequence, and their ability to develop in stages was registered in our consciousness as the concept of time.

---

* Two earlier volumes, viz. *Molecular Physics* and *Electricity and Magnetism*, were published by Mir Publishers in 1985 and 1986 respectively. The last volume in this series. *Mechanics and the Theory of Relativity*, is under preparation.

The above-mentioned basic properties of classical mechanics and the concepts of space and time formed the basis on which the development of classical physics has taken place right up to the present time. The interrelations between these concepts, and the concepts themselves, were modified in the course of advancement of our knowledge about them, but they have always served as the base for classical physics. The most important of these results led to the establishment of a continuous relation between space, time, material and motion. In the philosophical aspect, these ideas are described by the discipline of dialectical materialism. According to the teachings of dialectical materialism, space and time are forms of existence of matter and hence cannot be conceived without matter. Motion is assumed to be the way of existence of matter. Matter, space, time and motion are always linked inseparably.

The Aristotelian mechanics was based on ideas of general nature to describe the mechanical motion of rigid bodies. These ideas fully retained their significance in the Newtonian mechanics, although Aristotle's theory of motion was replaced by Newtonian theory after a reign of two thousand years. Aristotle believed that all types of motion of material bodies can be divided into two categories, viz. "natural" and "forced". "Natural motion" occurs on its own and does not involve the application of any force. It is meaningless to look for the reason behind the "natural motion" of a material body. To be more precise, the question as to why a certain natural motion takes place is always answered by the readymade undisputable statement that the motion is natural and occurs in a particular way and in no other way, without the application of any external influence. "Forced motion" does not occur on its own and is always initiated by some external influence which is described with the help of the concept of force. The answer to the question "why does a certain forced motion occur?" is that the body is subjected to a force under whose influence it moves in a particular way. Aristotle considered the upward motion of light objects, the downward motion of heavy objects, and the motion of heavenly bodies over the celestial sphere as forms of natural motion, and all other types of motion as forced. Note that if a body is at rest due to its inability to perform "natural motion", its state of rest is considered "forced". For example, if a body is at rest on a horizontal table, the absence of its motion in the vertical direction is "forced", and is due to an appropriate force acting in the vertical direction. Similarly, the absence of its motion in the horizontal direction is also due to the absence of an appropriate force acting along the horizontal. This shows that the law of motion cannot form the basis for defining the force, although force can be determined from the law of motion. This remark also fully refers to the attempts at using Newton's second law of motion as the definition of force. In the Aristotelian mechanics, force is responsible for the velocity of a body, while the concept of acceleration does not exist.

In the Newtonian mechanics, the "natural motion" in the sense of Aristotle's theory is the uniform rectilinear motion of a material point. The formulation of Newton's first law of motion sets the conditions under which this "natural (inertial) motion" is realized. It enables us to choose a system of coordinates in which such a "natural motion" can take place. According to Newton's second law, force does not impart a velocity to a material point, but is responsible for its acceleration, and that too only in the system of coordinates in which the velocity of motion is constant in the absence of the applied force, i.e. the motion is "natural". As in the Aristotelian mechanics, force takes into account the influence of external agencies on the motion of a body. Material bodies are the source of force, and hence force is a quantitative measure of the interaction of material bodies. Newton's third law states that the force with which one of the interacting bodies acts on the other is equal and opposite to the force with which this second body acts on the first. This aspect of forces is not considered in the Aristotelian mechanics.

Thus, Newton's theory of motion is a radically new step in comparison with Aristotle's theory. Among the basic new points of the theory, we can mention all the questions concerning the introduction of a new system of coordinates, including the principle of relativity, the properties of interacting bodies, a new equation of motion and the subsequent development of the concepts of space and time. However, the basic concepts of the Aristotelian mechanics and the approach to the problem of motion in the Newtonian mechanics have not undergone any significant changes.

The next important step in this direction was the creation of the special theory of relativity. Its revolutionary nature was reflected in its new approach to the problem of space and time. As a consequence of this, the inseparable link between space, time and motion (which was pointed out in dialectical materialism a long time before the creation of the theory of relativity) became a fundamental concept of the physical theory. As regards its content, however, the special theory of relativity belongs entirely to classical physics. Following the creation of the general theory of relativity, the inseparable connection between space, time, motion and matter became the basic concept of the most general physical theory. Like the special theory of relativity, the general theory also belongs to the classical physics in its content. The theory in which the laws of general or special theory of relativity play a significant role is called the relativistic theory. If the quantum regularities can be neglected in these theories, they belong entirely to the classical physics.

The Newtonian mechanics of a point served as the basis for the construction of the mechanics of an aggregate of points comprising a material body or a medium. If the motion of individual points can be described in accordance with Newton's laws, the theory pertains completely to the classical physics. In many cases, the concept of a continuous medium is employed in the mechanics of a body or a medium. According to this concept, mass is assumed to be uniformly "smeared" over the space, and the motion of a mass element in an infinitely small volume is described by the laws of mechanics of a point. Under this condition, the mechanics of continuous media also belongs to classical physics. In this connection, it is appropriate to make the following remark concerning the mechanics of a rigid body. The equations of motion of a rigid body include three equations for the coordinates of the centre of mass and three momentum equations. Strictly speaking, these six equations cannot be derived just from the three Newton's laws of motion for a material point. Their derivation also requires the additional assumption that the forces of interaction of material points constituting a rigid body are central in nature. However, even under this assumption the mechanics of a rigid body still belongs to the classical physics.

For analyzing the motion of a system of many real points each of which moves in accordance with Newton's laws of motion, a dynamic description of the system is impossible from a technical point of view, unsuitable from the theoretical viewpoint, and of no use in practice. For many-particle systems, new laws of motion have to be used due to the existence of a large number of particles in the system. These laws are called statistical laws. Statistical physics, whose dynamic laws are also Newton's laws of motion, belongs to classical physics and is often called classical statistical physics. It must be emphasized, however, that a complete and consistent substantiation of this branch is possible only with the help of quantum theory.

In all branches of classical physics considered above, the object of investigation was matter in the form of a substance. Another form of matter, which has been investigated successfully in physics, is called the field form. The electric and magnetic effects were discovered a long time ago, but the theory of these phenomena was developed rather slowly and culminated in Maxwell's theory only in 1860's. This was followed by the discovery of electromagnetic waves which can exist independently of the charges and currents generating them. This served as an experimental confirmation of the independent existence of the electromagnetic field and led to the concept of electromagnetic field as a form of matter. The motion of this form of matter is described by Maxwell's equations. These equations represent the law of motion of electromagnetic field and describe its formation by moving charges. The action of an electromagnetic field on charges whose carriers are material particles (corpuscles) is described by the Lorentz force. Maxwell's equations are derived from the basic concepts of strength and induction of electromagnetic field at points which vary in space with time, and the concept of electromagnetic field produced by a charge moving along a definite trajectory in the same way as a material point, and acting on this charge. This means that the theory based on Maxwell's equations belongs to classical physics. It is a relativistically invariant theory and pertains entirely to relativistic classical physics.

After the discovery of the field form of matter as electromagnetic waves and the creation of the electromagnetic theory of light, it became really possible to formulate the laws of mutual transformation of matter in the field or corpuscular form or, in other words, to solve the problem of interconversion of radiation and matter. It appeared that the problem should be solved within the framework of classical physics since each of these forms of matter can be easily described by an appropriate classical theory. The first indication that classical physics is not in a position to explain the interrelation between these two forms of matter was obtained during an analysis of blackbody radiation, where it was found necessary to admit the discreteness of the acts of emission. This was followed by the discovery of corpuscular properties of radiation and the wave properties of electrons and other particles. These discoveries confirmed that there is no barrier between the corpuscular and field forms of matter, that these forms overlap and exist in a dialectical unity. Experimental investigation and analysis of this dialectical unity led to a complete revision of the basic concepts of classical physics and to the creation of quantum theory.

The philosophical categories of dialectical materialism are realized quite concretely and implemented in classical physics, while the methodological principles of physical investigations had a profound impact on the analysis of gnosiological problems. Classical physics contains a complete and all-round description of all the aspects of interrelation between science and philosophy. This circumstance has a great philosophical significance and has been aptly reflected in all the four volumes of this course of physics. The dialectical unity of space, time, motion and matter has been described fully and has considerably influenced the style and content of this course. For example, the manner in which the concepts of space, time and motion are expounded, as is often done, in the framework of kinematics without establishing an intrinsic link between them is unacceptable while the relation of these concepts with that of matter is first discussed only in dynamics after the introduction of the concept of mass. This discontinuity contradicts the very essence of space and time as forms of existence of matter, and of motion as a manner of this existence. We have removed this discontinuity by introducing the reader at the very beginning of the course in physical kinematics to the ideas of the theory of relativity which gives a very clear account of the concepts of dialectical materialism about the inseparable link between space, time, motion and material*. This dialectical unity is followed up and refined in the subsequent parts of this course. The questions concerning the general relation between phenomena, indestructibility of matter and motion, causality and determinism, interpretation of laws as expressions connecting various phenomena, etc. are fully described in the course of classical physics. In short, in classical physics the realization of general philosophical categories of dialectical materialism and statements of Marxist-Leninist philosophy in the specific forms of knowledge is so full and perfect that the problem about the nature of inadequacy of this special knowledge and about the content of incompleteness of the unity between the specific knowledge and general philosophical and gnosiological categories is of immense importance. The importance of this problem stems from the fact that the inadequacy of specific knowledge and of its unity with general philosophical and gnosiological categories is the only source and driving force behind the development of both the specific knowledge and the general philosophical and gnosiological categories. In classical physics, this inadequacy appears only in a potential form and does not serve as a real negation of the completeness. The completeness of knowledge attained in classical physics and its unity with general philosophical and gnosiological categories is negated only in quantum physics. According to the dialectics of negation, not only this leads to a further development of science, but also stimulates the analysis of general philosophical and gnosiological problems.

---

* See *Mechanics and Theory of Relativity* by the same author (to be published) – *Tr.*

Subject Index

Aberration,
  chromatic, 174
  geometrical, 173,
  longitudinal, 175
  optical, 173
  spherical, 175
  transverse, 175
accommodation, 182
achromatization, 178
amplification of light, parametric, 427
amplifier(s), optical, 391
analyzer, 347
angle,
  Brewster, 132
  coherence, 209
anisotropy,
  artificial, 358
  optical, 331
  total, 331
approximation,
  Fresnel, 276
  Kirchhoff, 273
  optical, 274
  paraxial, 159
  ray, 155
averaging, 52
  of harmonic functions, 53
axis, optical, 340

Beam, Gaussian, 36
beats, 44
biprism, Fresnel's, 213
birefringence, 344
blackbody, 384
broadening, 86
  collision, 87
  Doppler, 88
  nonuniform, 86
  uniform, 86

Camera,
  photographic, 183
  Schmidt, 185
candela, 61
cell,
  Kerr, 360
  Pockels, 361
coefficient,
  Einstein, 389
  reflection, 140
  scattering, 372
  transmission, of lens, 140
coherence
  partial, 239
  space, 209
  spatial, 210
  time, 209
coma, 176
compensator, 352
component,
  Brillouin, 375
  undisplaced, 375
condition,
  Abbe sine, 179
  amplification, 392
  boundary, 121
  cyclicity, 399
  Petzval, 177
  quasi-monochromaticity, 86
  radiation, 271
  saturation, 393
  steady-state lasing, 396
  of vector spatial synchronism, 424
constant,
  Boltzmann, 19
  dielectric, 331
  Gauss, 167
  Kerr, 359
  Kundt, 358

- Planck, 16
  - electric, 23
  - electrodynamic, 22
  - magnetic, 23
- Stefan-Boltzmann, 387
- Verdet, 357

correlation interval, 108
crystal(s),
- biaxial, 341
- negative, 344
- positive, 344
- uniaxial, 341

curve,
- caustic, 176
- dispersion, 115

Degree of polarization, 134
density,
- energy flux, 35
- momentum, 37
- number, 384
- radiant flux, 57
- radiation, 381
  - equilibrium, 382
  - spectral, 382
- spectral, of radiation, 58
- volume, 47

depth, penetration, 137, 143
detection, synchronous, 315
diagram,
- aperture, 181
- field-of-view, 182

diaphragming, 180
diffraction, 97
- Fraunhofer, 277
- Fresnel, 266, 276, 294

diopter, 170
dispersion, 94, 112
- anomalous, 116
- normal, 115

distortion, 177
distribution, Gaussian, 37

Effect,
- Brewster, 132
- Doppler, 33
  - longitudinal, 33
  - transverse, 33
- Faraday, 357
- Kerr, 360
- Pockels, 361

efficacy,
- spectral luminous, 63
  - relative, 64

Einstein-Podolskii-Rosen paradox, 193
ellipsoid,
- of group velocities, 340
- index, 341

emittance,
- luminous, 62
- radiant, 59

energy,
- of photon, 16
- of wave train, 41

equation(s)
- Abbe, 180
- eikonal, 153
- Fresnel, 335
- index ellipsoid, 341
- lens, 168
  - in Gaussian form, 169
  - in Newton's form, 169
- Maxwell, 23
- wave, 23

experiments,
- Abbe-Porter, 314
- Brown-Twiss, 243

exposure, 63

Filter, interference, 224
focus, of optical system, 167
force,
- Coulomb repulsive, 39
- radiometric, 38

formula,
- Brillouin, 375
- Fresnel, 127
- Green, 269
- Rayleigh-Jeans, 385
- Wien, 386

Fourier optics, 299
frequency(ies),
- beat, 45
- carrier, 91
- circular (cyclic), 15
- modulated, 92
- negative, 78
- sideband, 91

fringe(s),
- uniform inclination, 228
- uniform thickness, 230

function,
- autocorrelation, 105
- coherence, 241

function,
  coherence,
    mutual, 239
  nonperiodic, 72
  normalized, 241
  periodic, 72
  position, 72
  self-coherence, 241
  spectral, 104
fusion, laser, 39

Generators of light, parametric, 427
glass, magnifying, 183
grating,
  amplitude-and-phase, 289
  diffraction, 283
  phase, 288
  reflection, 287

Hologram, 316
  computer, 327
  digital, 327
  thick, 323
holography, 315
Huygens construction, 344

Illuminance, 63
  radian, 60
image, optical, 165
integral,
  Fourier, 71
    in complex form, 72
    in real form, 72
  Fresnel, 295
interference, 189
  of monochromatic waves, 191
  of nonmonochromatic light, 196
  two-beam, 205
  in white light, 206
interferogram, 198
interferometer,
  Fabry-Perot, 216
  Mach-Zehnder, 201
  Michelson, 191
  Rayleigh, 212
  scanning, 223
  stellar, 211
  Twyman-Green, 202
  Young, 205
interferometry, holographic, 326
invariance,
  phase, 32
  of plane wave, 31
isomerism, optical, 357

Kronecker delta, 332

Laser(s),
  CO_2,
    closed-volume, 409
    CW-mode, 410
  continuous wave, 397
  dye, 410
  gasdynamic, 410
  helium-neon, 408
  pulsed, 397
  ruby, 406
  schematic diagram of, 394
  T-, 410
law,
  Biot, 355
  Bragg, 323
  Brewster, 132
  of energy conservation, 141
  Lambert, 62
  Malus, 346
  of reflection, 122
  of refraction, 122
  Snell, 124
length,
  coherence, 101
  focal, 167
  optical, 226
  self-focussing, 428
lens,
  Billet, 213
  oil-immersion, 178
  thin, 169
light,
  completely polarized, 246
  Gaussian, 101
  natural, 245
  random, 98
light pressure, 37
line, spectral,
  anti-Stokes, 375
  Gaussian, 89
  Lorentz, 83
  Stokes, 375
luminance, 62
luminous flux, 61

Mask, 328
matrix,

  coherence, 244
  refraction, 161
  transfer, 161
  transformation, 166
medium(a)
  active, 394
  anisotropic, 330
  dielectric, 331
method,
  dark-field illumination, 309
  Fresnel zone, 263
  phase-contrast, 310
  Q-switching, 397
  of thick holograms, 323
Michelson's echelon, 225
microscope, 183
mirror,
  Fresnel, 214
  Lloyd, 213
  multilayer dielectric, 237
mode(s),
  axial, 400
  radiation, 398
  side, 401
  synchronization of, 403
modulation,
  amplitude, 91
  frequency, 92
  phase, 92
momentum, of wave train, 42

Operator,
  D'Alembert, 24

Permittivity,
  complex, 142
  relative 111
photon(s), 17
  noise, 20
  solar-radiation, 19
  thermal, 20
plane,
  focal, 167
  meridional, 177
  principal, 166
  sagittal, 177
plate,
  half-wave, 350
  Lummer-Gehrcke, 225
  quater-wave, 350
  wave, 350
polarization, 49
  circular, 50
  degree of, 134
  elliptical, 50
  of light, 141
  linear, 49, 319, 413
  nonlinear, 319, 413
polarizer, 347
polaroid, 347
polychroism, 347
power,
  absorptive, 382
  emissive, 57
  resolving, 219
  rotatory, 354
prism,
  birefringent, 347
  Nicol, 347
  polarizing, 347
principle,
  Fermat, 155
  Huygens, 205
principle,
  Huygens-Fresnel, 263
  tautochronism, 156
projectors, optical, 185
process(es), random, 103
pulse(s), 97
pupil,
  entrance of, 182
  exit of, 182

Q-factor, 395
quantities,
  photometric, 60
  radiant, 57

Radiant intensity, 58
radiation,
  blackbody, 381
  laser, 398
radiator,
  classical model of, 79
  elementary, 58
range,
  optical, 16
  visible, 16
relation, Fresnel-Kirchhoff, 273
ray,
  extraordinary, 344
  incident, 123
  of light, 155
ray,

  ordinary, 344
  reflected, 123
  refracted, 123
radiance, 59
  spectral, 58
refractometer,
  Jamin, 203
ring(s),
  interference, 218
  Newton, 231
region, dispersion, 222

Scatterer, elementary, model of, 366
scattering,
  Brillouin, 375
    in solids, 376
  of light, 365
  Mie, 372
  multiple, 366
  nature of, 365
  Raman, 376
  Rayleigh, 368
  types of, 365
self-focussing,
  of beam, 428
  of light, 427
series, Fourier,
  in complex form, 72
  in real form, 71
shape,
  of absorption line, 84
  of composite emission line, 90
  of emission line, 85
source(s), of anisotropy, 331
speckles, laser, 405
spectrum,
  amplitude, 73
  continuous, 73
  phase, 73
  power, 103
  of pulses, 74
  rectangular, 74
  sawtooth, 75
  single rectangular, 75
spiral, Cornu, 267, 295
spot, Poisson, 266
superposition,
  of electromagnetic waves, 43
  of linearly polarized waves, 49
surface,
  caustic, 176
  Lambert, 62
  ray, 342
susceptibility, 113
  complex, 413
  dielectric, 413
  nonlinear, 415
  second-order, 416
synchronization of modes, 403
synchronism, spatial, 423
system, Maksutov, 185

Telescope, 183
tensor, permittivity, 332
theorem,
  Helmholtz reciprocity, 274
  Helmholtz-Kirchhoff, 269
  Parseval, 78
  Plancherel, 79
  Van-Zittert-Zernike, 253
  Wiener-Khintchine, 107
threshold, lasing, 395
time,
  coherence, 101
  emission, 84
  resolution, 99
time lag, 431
transform, Fourier, 73
transition,
  induced, 389
  spontaneous, 389

Vector,
  Poynting, 35
  wave, 28
velocity,
  group, 95
  phase, 338
  ray, 338
visibility, 199f
vision, night, 21

Wave(s),
  coherent, 190
  convergent, 27
  divergent, 27
  electromagnetic, 22
  harmonic, 27
  incident, 121
  linear polarization, 417
  modulated, 91
  monochromatic, 111
  plane, 24
    harmonic, 27
  polarized, 50

  circularly, 50
  elliptically, 50
  partially, 50
quasi-monochromatic, 86
quasi-plane, 97
reference, 316
reflected, 121
refracted, 121
signal, 316
spherical, 26
  standing, 45
  travelling, 27
wave front, 155
wave number, 112
wave packet, 94
width, coherence, 209

Zone,
  Fresnel, 263
  zeroth, 264

TO THE READER

MIR PUBLISHERS would be grateful for your comments on the content, translation and design of this book.

We would also be pleased to receive any other suggestions you may wish to make.

Our address is:

Mir Publishers 2 Pervy Rizhsky Pereulok I-110, GSP, Moscow, 129820 USSR